Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και
Υπολογιστών

Πρόβλεψη Καθυστέρησης Πτήσεων με Τεχνικές Μηχανικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΚΩΝΣΤΑΝΤΙΝΟΣ ΑΛΕΞΑΝΔΡΟΥ

Επιβλέπων : Ανδρέας - Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2021

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και
Υπολογιστών

Πρόβλεψη Καθυστέρησης Πτήσεων με Τεχνικές Μηχανικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΚΩΝΣΤΑΝΤΙΝΟΣ ΑΛΕΞΑΝΔΡΟΥ

Επιβλέπων : Ανδρέας - Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 14η Ιουλίου 2021.

.....	.....	.....
Ανδρέας - Γεώργιος Σταφυλοπάτης Καθηγητής Ε.Μ.Π.	Γεώργιος Στάμου Αν. Καθηγητής Ε.Μ.Π.	Στέφανος Κόλλιας Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2021

Κωνσταντίνος Αλεξάνδρου

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Κάθε χρόνο οι καθυστερήσεις πτήσεων οδηγούν σε απώλειες δισεκατομμυρίων δολαρίων ως επακόλουθο της καθυστέρησης παροχής υπηρεσιών, της αργοπορίας των υλικών αγαθών να φτάσουν στον προορισμό τους και του προσωπικού κόστους κάθε επιβάτη. Το πρόβλημα της αντιμετώπισης των καθυστερήσεων αποτελεί ζήτημα εξαιρετικής δυσκολίας, όπως είναι βέβαια και η ικανότητα αναμονής και πρόβλεψης των εν λόγω καθυστερήσεων. Στην παρούσα εργασία, χρησιμοποιώντας τεχνικές επιβλεπόμενης μηχανικής μάθησης θα μελετηθεί η δυνατότητα κατηγοριοποίησης των πτήσεων σε πτήσεις με καθυστέρηση ή όχι με την μοντελοποίηση του προβλήματος ως ένα πρόβλημα δυαδικής ταξινόμησης. Μέσω βιβλιογραφικής έρευνας και πειραματικών δοκιμών θα ταυτοποιηθούν τα χαρακτηριστικά που επιτρέπουν στα μοντέλα μηχανικής μάθησης να εκπαιδευτούν και να προβλέψουν ενδεχομένη καθυστέρηση. Για την ταυτοποίηση των κατάλληλων χαρακτηριστικών, δοκιμάστηκε η ενσωμάτωση στα δεδομένα πτήσεων, πληθώρας μετεωρολογικών χαρακτηριστικών και χαρακτηριστικών ενός αεροπορικού δικτύου. Ακόμη, θα εξεταστεί και θα συγκριθεί η επίδοση απλών ταξινομητών, τεχνητών νευρωνικών δικτύων και μεθόδων Ensemble για την κατηγοριοποίηση των πτήσεων. Από τα τελικά πειραματικά αποτελέσματα, προκύπτει πως πράγματι η μηχανική μάθηση μπορεί να εφαρμοστεί επιτυχώς για τη δημιουργία ενός συστήματος πρόβλεψης καθυστέρησης πτήσεων.

Λέξεις κλειδιά

Μηχανική μάθηση, επιβλεπόμενη μάθηση, καθυστέρηση πτήσεων, ενίσχυση κλίσης, τεχνητά νευρωνικά δίκτυα, δυαδική ταξινόμηση, λογιστική παλινδρόμηση, XGBoost, Τυχαία δάση

Abstract

Every year flight delays cost billions of dollars as a result of services being delayed, consumer goods arriving late and passengers not arriving on time at their destinations. The problem of dealing with flight delays is an extremely challenging one, as is, of course, accurately anticipating and predicting these delays. With the use of supervised machine learning techniques, we will attempt to classify flights in two classes, “On-time” and “Delayed”, thus modelling the problem as a binary classification one. In the present thesis, through scientific literature research and experimental tests, features are identified that allow the machine learning models to be trained and to predict possible delays. In order to identify the appropriate features, a variety of meteorological and airspace network characteristics were tested, then chosen and integrated into our datasets. The performance of simple classifiers, Artificial Neural Networks and Ensemble algorithms is examined and compared on our flight data. After the experimentation we were able to conclude that machine learning techniques can in fact be implemented in designing a model that can accurately predict flight delays.

Key words

Machine Learning, Supervised Learning, flight delay, gradient boosting, Artificial Neural Networks, binary classification, logistic regression, XGBoost, Random Forest

Ευχαριστίες

Ξεκινώντας, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα μου, τον Καθηγητή κ. Ανδρέα - Γεώργιο Σταφυλοπάτη για την εμπιστοσύνη και την ευκαιρία που μου έδωσε να ασχοληθώ με την παρούσα εργασία. Εν συνεχεία, δεν θα μπορούσα να μην ευχαριστήσω τον κ. Γεώργιο Σιόλα τόσο για την καθοδήγηση του όσο και για την εξαιρετικής σημασίας βοήθεια και στήριξη που μου παρείχε καθ' όλη τη διάρκεια εκπόνησης της παρούσας εργασίας. Οι συμβουλές και οι υποδείξεις του κ. Σιόλα ήταν καθοριστικές για την τελική μορφή της εργασίας και χωρίς την πολύτιμη συμβολή του, δεν θα έφτανα σε αυτό το σημείο. Από τον κύκλο των ευχαριστιών μου, δε θα μπορούσαν να λείπουν εναπομείναντα δύο μέλη της Τριμελούς Εξεταστικής Επιτροπής, ο κ. Γεώργιος Στάμου, Αναπληρωτής Καθηγητής του Ε.Μ.Π., και ο κ. Στέφανος Κόλλιας, Καθηγητής Ε.Μ.Π., τόσο γιατί δέχτηκαν να απαρτίσουν την Επιτροπή της διπλωματικής αυτής εργασίας, όσο και για το ενδιαφέρον που έδειξαν.

Θα ήθελα ακόμη, να ευχαριστήσω όλους τους κοντινούς μου ανθρώπους που στάθηκαν δίπλα μου σε αυτό το μεγάλο ταξίδι και με βοήθησαν, ο καθένας με τον τρόπο του. Τέλος, ένα μεγάλο ευχαριστώ οφείλω στην αγαπημένη μου οικογένεια και ειδικά στους γονείς μου για την στήριξη και την εμπιστοσύνη τους, όντας πάντα στο πλευρό μου όλα αυτά τα χρόνια.

Κωνσταντίνος Αλεξάνδρου,

Αθήνα, 6η Ιουλίου 2021

Περιεχόμενα

Κατάλογος πινάκων

Κατάλογος σχημάτων

Κεφάλαιο 1

ΕΙΣΑΓΩΓΗ

1.1 Το πρόβλημα της Καθυστέρησης Πτήσεων

Οι μοντέρνες κοινωνίες στηρίζονται πάνω στην ευκολία μετακίνησης υλικών αγαθών, ανθρώπων και υπηρεσιών. Εφοδιαστικές αλυσίδες διασχίζουν ολόκληρες χώρες και ηπείρους ενώ ταυτόχρονα οι άνθρωποι ταξιδεύουν όλο και περισσότερο για επαγγελματικούς και προσωπικούς σκοπούς. Η αύξηση ζήτησης εύκολης μετακίνησης οδηγεί σε ψηλή εξάρτηση από συστήματα μεταφορών και τις υπηρεσίες που παρέχουν: κι όταν αυτά τα συστήματα αντιμετωπίζουν περιορισμούς, αυτό μεταφράζεται σε μεγάλο χρηματικό κόστος [1], [2].

Ένας τομέας που πλήττεται συχνά από καθυστερήσεις είναι εκείνος της αεροπλοΐας. Σύμφωνα με το Bureau of Transportation Statistics των ΗΠΑ, μεταξύ Ιανουαρίου 2012 και Δεκεμβρίου 2019, περίπου 20% των εσωτερικών πτήσεων εντός των Ηνωμένων Πολιτειών της Αμερικής είχαν καθυστερήσεις ή οδηγήθηκαν σε ακύρωση [3]. Γενικά, καθυστερήσεις προκύπτουν όταν η αλληλεπίδραση μεταξύ των οντοτήτων που συμβάλλουν στις αερομεταφορές (εταιρείες αερομεταφορών, αεροδρόμια και πύργοι ελέγχου) και των εξωτερικών παραγόντων (ακραίες καιρικές συνθήκες, απεργίες, απρόσμενα γεγονότα) είναι τέτοια ώστε να οδηγήσει σε συμφόρηση [4].

Οι καθυστερήσεις στις μεταφορές δημιουργούν κόστος: τόσο για την οικονομία της χώρας αυτή καθαυτή, με την καθυστέρηση υπηρεσιών και παράδοσης υλικών αγαθών στον προορισμό τους, όσο και εξατομικευμένα για τον εκάστοτε ταξιδιώτη. Σημαντική ερευνητικό αποτέλεσμα προέκυψε από τους Ball και Barnhart και αφορούσε στην μοντελοποίηση του αντίκτυπου στο χρηματικό κόστος μιας καθυστέρησης [1]. Σύμφωνα με στοιχεία που παρέχονται από το Federal Aviation Administration (FAA) τα κόστη καθυστερήσεων τα τελευταία χρόνια ανέρχονται σε δεκάδες δισεκατομμύρια δολάρια. Συγκεκριμένα, το 2016 το συνολικό κόστος από τις καθυστερήσεις ανήλθε στα 23.7 δισεκατομμύρια δολάρια, ενώ το 2019 το αντίστοιχο ποσό ήταν 33 δισεκατομμύρια δολάρια [5]. Για τις αεροπορικές εταιρείες μόνο, το κόστος το 2019 ανήρθε στα $8.3 δισεκατομμύρια. Τα εν λόγω κόστη συμπεριλαμβάνουν και αυξημένα λειτουργικά κόστη, όπως επιπλέον πληρώματα εδάφους και

αέρος, καύσιμα, κόστη συντήρησης κ.α., τα οποία προκύπτουν από απρόβλεπτες καθυστερήσεις. Τα κόστη για τους επιβάτες ανήρθαν στα 18.1 δις. δολάρια και αντιστοιχούν σε ακυρώσεις πτήσεων, απώλεια πτήσεων ανταπόκρισης και καθυστερήσεις που έχουν αλυσιδωτές συνέπειες στα σχέδια του επιβάτη.

	2016	2017	2018	2019
Airlines	5.6	6.4	7.7	8.3
Passengers	13.3	14.8	16.4	18.1
Lost Demand	1.8	2.0	2.2	2.4
Indirect	3.0	3.4	3.9	4.2
Total	**23.7**	**26.6**	**30.2**	**33.0**

Πίνακας 1: *Total Cost of Delay in the U.S. (dollars, billion), FAA*

Όπως είναι φανερό, τα κόστη που προκύπτουν από αυτό το συχνό φαινόμενο είναι υπέρογκα και επιβαρύνουν όλους τους εμπλεκόμενους. Οι άμεσα ενδιαφερόμενοι αλλά και η επιστημονική κοινότητα καλούνται να βρουν τρόπους προκειμένου να μετριασθεί η ζημία που προκαλείται από τις απρόβλεπτες καθυστερήσεις.

1.2 Αντικείμενο της Διπλωματικής

Στην παρούσα διπλωματική εργασία θα διατυπωθεί το πρόβλημα της καθυστέρησης πτήσεων και θα αναζητηθούν τρόποι για την αποτελεσματική πρόβλεψη τους. Το πρόβλημα αυτό θεωρείται ανοιχτό από την επιστημονική κοινότητα και διενεργούνται ποικίλες και εξαιρετικής σημασίας επιστημονικές έρευνες που το προσεγγίζουν. Στην παρούσα εργασία, θα επιχειρήσουμε χρήσει τεχνικών Μηχανικής Μάθησης να προβλέψουμε εάν μια συγκεκριμένη πτήση θα καθυστερήσει ή όχι.

Η εργασία αφορά το National Aviation System (NAS) των ΗΠΑ και βασίζεται πάνω σε ανοιχτά δεδομένα που παρέχονται από κυβερνητικούς οργανισμούς στο διαδίκτυο. Η σκοπιά γύρω από την οποία προσεγγίζεται το εν λόγω πρόβλημα, είναι αυτή της μελέτης των παραγόντων που επηρεάζουν την ομαλή διεξαγωγή μιας προγραμματισμένης πτήσης και ο τρόπος με τον οποίο αυτοί οι παράγοντες αν αξιοποιηθούν κατάλληλα και χρησιμοποιηθούν

σε μια διαδικασία εκπαίδευσης μοντέλων ΜΜ μπορούν να βοηθήσουν στην πρόβλεψη των καθυστερήσεων.

Όπως ορίζεται από την FAA, μία πτήση θεωρείται ότι έχει καθυστερήσει εάν έχει παρέλθει χρονικό διάστημα μεγαλύτερο των 15 λεπτών από την προγραμματισμένη άφιξη της πτήσης. Βάσει αυτού του κανόνα λοιπόν, θα ταξινομήσουμε τις πτήσεις στην ανάλογη κατηγορία.

Ως αεροδρόμιο ενδιαφέροντος για την παρούσα εργασία επιλέγηκε το John F. Kennedy International Airport (JFK) στην Νέα Υόρκη, καθώς είναι ένα από τα μεγαλύτερα στην χώρα αλλά και παγκοσμίως. Δεδομένου του τεράστιου μεγέθους του Εθνικού Αεροπορικού Δικτύου των ΗΠΑ θα εξεταστεί ένα μικρό υποσύνολο των αεροδρομίων που εκτελούν εσωτερικές πτήσεις προς το JFK, το οποίο αποτελείται από τρία μεγάλα αεροδρόμια των ΗΠΑ σε αρκετά διαφορετικές ως προς το κλίμα περιοχές των ΗΠΑ. Πιο συγκεκριμένα, θα μελετηθούν το San Francisco International (SFO), το Dallas Fort Worth International (DFW) και το Chicago O'Hare International Airport (ORD).

Σχήμα 1: *Χάρτης ΗΠΑ με αεροδρόμια ενδιαφέροντος*

Εν προκειμένω, θα μελετήσουμε και θα προτείνουμε τι είδους πληροφορία είναι αναγκαία για την σωστή πρόβλεψη μιας πτήσης αλλά και θα αποφανθούμε μέσα από πληθώρα μοντέλων ΜΜ και τεχνικών προεπεξεργασίας των δεδομένων, ποια είναι τα καταλληλότερα εργαλεία για την αντιμετώπιση του προβλήματος.

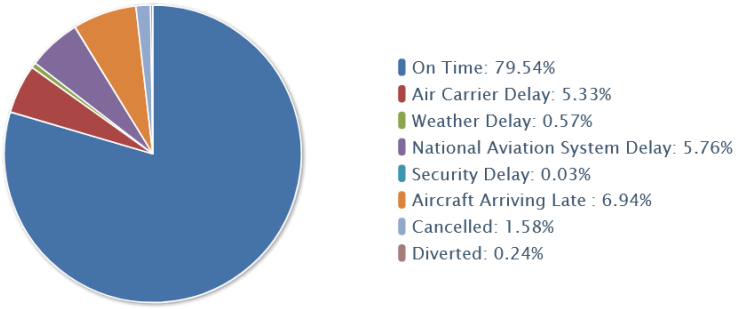
1.3 Γενική Προσέγγιση του Προβλήματος

Ως βάση για τη μελέτη του προβλήματος χρησιμοποιήθηκαν δεδομένα πτήσεων που συγκεντρώθηκαν από το BTS για την περίοδο 01/01/2012 – 31/12/2019. Ωστόσο, θα

περιοριστούμε στην μελέτη αυτών που αφορούν το αεροδρόμιο JFK ως τελικό προορισμό. Γνωρίζοντας όμως ότι η ομαλή διεξαγωγή μιας πτήσης επηρεάζεται από πληθώρα εξωγενών παραγόντων, προκύπτει το συμπέρασμα πως δεν είναι αρκετή μονάχα η μελέτη αρκετά μεγάλου συνόλου δεδομένων για την εύρεση μοτίβων στα δεδομένα. Ως εκ τούτου, τα δεδομένα για τον σκοπό της παρούσας εργασίας συνδυάστηκαν με μετεωρολογικά δεδομένα που αφορούν σταθμούς κοντά στα αεροδρόμια στα οποία εστιάσαμε, καθώς και με πληροφορία που εξήχθη από δεδομένα που αφορούν ολόκληρο το αεροπορικό δίκτυο, όπως ο όγκος των πτήσεων σε ημερήσια βάση καθώς και την ιχνηλάτηση των αεροσκαφών που εκτελούν τις πτήσεις άμεσου ενδιαφέροντος. Παρέχοντας την περισσότερη δυνατή πληροφορία, ειδικά μορφοποιημένη στα μοντέλα μηχανικής μάθησης που κατασκευάστηκαν, στόχο έχουμε την επίτευξη υψηλής ορθότητας των προβλέψεων μας που αφορούν καθυστέρηση πτήσεων.

1.4 Παράγοντες που συμβάλλουν στην καθυστέρηση

Όπως προαναφέρθηκε, μια πτήση μπορεί να επηρεαστεί από πληθώρα εξωγενών παραγόντων. Για κάποιους από τους παράγοντες αυτούς μπορούν να ληφθούν προληπτικά μέτρα, όπως η βελτιστοποίηση της διαδικασίας προετοιμασίας μιας πτήσης από τον αερομεταφορέα, ή η αποδοτικότερη ολοκλήρωση των λειτουργιών ενός αεροδρομίου από το προσωπικό κ.α. Υπάρχουν όμως και άλλοι παράγοντες που επηρεάζουν τη δυνατότητα ομαλής διεξαγωγής μιας πτήσης οι οποίοι δεν μπορούν να ελεγχθούν. Ένας τέτοιος παράγοντας είναι οι καιρικές συνθήκες. Σύμφωνα με το BTS για την περίοδο 2012-2019 το **18.64%** των πτήσεων έφτασε στον προορισμό του με καθυστέρηση. Το **28,61%** εξ αυτών οφείλεται σε παράγοντες που ο αερομεταφορέας μπορούσε να προλάβει το **37,25%** οφείλεται σε προηγούμενη καθυστέρηση του αεροσκάφους που θα εκτελέσει την πτήση και το **30,91%** οφείλεται σε καθυστέρηση που προκλήθηκε λόγω του Εθνικού Αεροπορικού Δικτύου NAS. Το **3,06%** των καθυστερήσεων οφείλεται σε **ακραίες καιρικές συνθήκες** ενώ μόλις το **0,17%** των καθυστερήσεων είναι ως αποτέλεσμα κάποιου περιστατικού ασφαλείας [3].

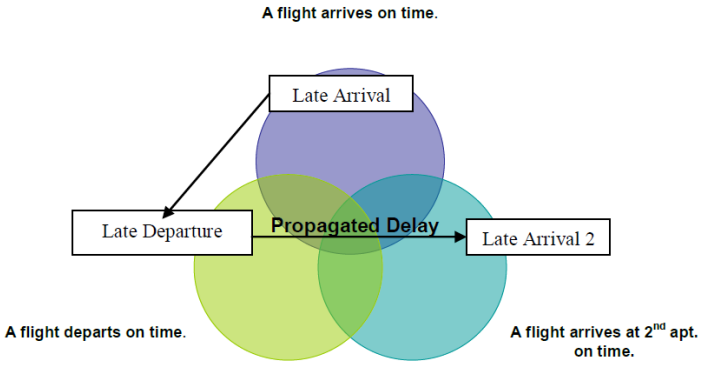


Σχήμα 2: *On-Time Arrival Performance (01/12-12/19), BTS*

Αργοπορία Αεροσκάφους

Είναι φανερό, πως ο κυριότερος παράγοντας για την πρόβλεψη καθυστέρησης μιας πτήσης είναι η γνώση του κατά πόσο το αεροσκάφος που εκτελεί ένα συγκεκριμένο δρομολόγιο έχει καθυστερήσει να φτάσει στο αεροδρόμιο αναχώρησης. Η διάδοση αυτής της καθυστέρησης μέσα στο δίκτυο αποτελεί ένα πολύ ενδιαφέρον ερευνητικό πρόβλημα και κατά καιρούς, έχει απασχολήσει αρκετούς ερευνητές. Σύμφωνα με την Kondo [6] για μια ακολουθία πτήσεων, προϋπόθεση διάδοσης της καθυστέρησης (propagation of delay) από τη μία στην άλλη είναι η ταυτόχρονη ισχύς τριών συνθηκών:

- Μία πτήση αφικνείται με καθυστέρηση
- Αναχωρεί με καθυστέρηση για το επόμενο τμήμα του ταξιδιού
- Αφικνείται με καθυστέρηση στον επόμενο προορισμό



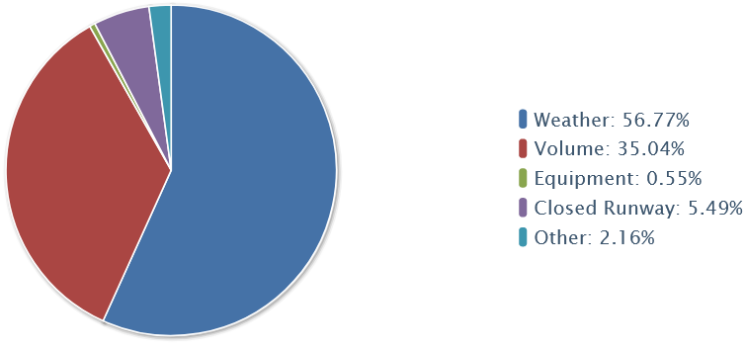
Σχήμα 3: *Venn diagram depicting the three condition for a propagated delay, Akira Kondo*

Στην ερευνητική εργασία με τίτλο “Delay Propagation and Multiplier” [6] μελετάται η διάδοση και ο πολλαπλασιασμός της καθυστέρησης μέσα στο αεροπορικό δίκτυο στις

επακόλουθες στάσεις που εκτελεί ένα αεροσκάφος μέσα στην ίδια μέρα. Η τεχνική που χρησιμοποιήθηκε από την ερευνητική ομάδα για την ιχνηλάτηση των αεροσκαφών είναι η ίδια με αυτήν που εφαρμόστηκε στην παρούσα εργασία, και αφορά τη χρήση του μοναδικού αριθμού (Tail Number) κάθε αεροσκάφους για την εύρεση των προηγούμενων ή επόμενων σταθμών του. Το φαινόμενο του delay propagation μελέτησαν επίσης με χρήση Bayesian Δικτυών οι Xu και Donohue. [7] και Wang και Schaefer. [8]

Καιρικές Συνθήκες

Οι αντίξοες καιρικές συνθήκες αποτελούν έναν από τους κυριότερους παράγοντες πρόκλησης καθυστέρησης. Μολονότι με μια πρώτη ματιά φαίνεται ότι η κακοκαιρία είναι υπαίτια μόνο για το **3,06%** των καθυστερήσεων, αυτό το νούμερο αφορά μόνο τις ακραίες καιρικές συνθήκες, δηλαδή αυτές οι οποίες αποτρέπουν εντελώς την απογείωση ενός αεροσκάφους και έτσι καθυστερούν την αναχώρηση και ως επακόλουθο την άφιξη του. Παρόλα αυτά υπάρχουν και τα καιρικά φαινόμενα που δεν κατηγοριοποιούνται ως ακραία αλλά εξακολουθούν να προκαλούν πολλές καθυστερήσεις στην αναχώρηση των πτήσεων διότι προσθέτουν επιπλέον βήματα στην προετοιμασία της απογείωσης, απαιτούν πιο προσεκτική και αργή μετακίνηση των αεροσκαφών στους χώρους του αεροδρομίου και καθυστερούν τις λειτουργίες των αερολιμένων. Οι συγκεκριμένες καθυστερήσεις όπως αυτές αναφέρονται παραπάνω, εμπίπτουν μέσα στην κατηγορία NAS (30,91% των καθυστερήσεων). Σε αυτή την κατηγορία οι καθυστερήσεις του καιρού αποτελούν την πλειοψηφία και συγκεκριμένα το **56.77%.** Σημειώνεται, πως δεύτερος σημαντικότερος παράγοντας είναι η συμφόρηση που προκαλείται από τον μεγάλο όγκο πτήσεων (35.04%).



Σχήμα 4: *Causes of National Aviation System Delays (07/18-12/19), BTS*

Προκειμένου να αποκτηθεί μια ακριβέστερη εικόνα του τρόπου με τον οποίο συγκεκριμένες συνθήκες επηρεάζουν την διεξαγωγή μιας πτήσης θα εξετάσουμε συγκεκριμένα

το πώς επηρεάζονται οι πτήσεις από τα στοιχεία της φύσης, κάτι που θα μας επιτρέψει στη συνέχεια να κάνουμε ορθή επιλογή των χαρακτηριστικών που θα χρειαστούμε στις προβλέψεις μας. Κάποιοι σημαντικοί παράγοντες είναι οι εξής:

- **Θερμοκρασία - Παγετός**

Οι πολύ χαμηλές θερμοκρασίες (κάτω των 0ºC) οδηγούν στη δημιουργία παγετού, τόσο στους αεροδιαδρόμους όσο και στα φτερά των αεροσκαφών. Το φαινόμενο δημιουργίας πάγου πάνω στο αεροσκάφος εμφανίζεται τόσο κατά τη διάρκεια της πτήσης όσο και κατά την παραμονή του αεροσκάφους στο έδαφος. Σύμφωνα με το National Transportation Safety Board, ο πάγος στα φτερά κατά την πτήση αποτελεί το 11% όλων των ατυχημάτων που οφείλονται στις καιρικές συνθήκες. Κατ'επέκταση, τα αεροσκάφη σε τέτοιες θερμοκρασίες πρέπει πριν την απογείωση να περνάνε διαδικασία από-παγοποίησης η οποία οδηγεί σε καθυστερήσεις, ειδικά στην περίπτωση που ο αριθμός των αεροσκαφών είναι μεγάλος και το εκπαιδευμένο πλήρωμα στο αεροδρόμιο περιορισμένο. Παρόμοια διαδικασία με ειδικά μηχανήματα πρέπει να γίνει και στους αεροδιάδρομους, με αποτέλεσμα οι απογειώσεις να είναι πιο αραιές και η αναμονή στο δίαυλο μεγαλύτερη. Για τον ίδιο λόγο τα αεροσκάφη κινούνται με μικρότερες ταχύτητες στο έδαφος για την αποφυγή ατυχημάτων. Πολλές πτήσεις αναγκάζονται να αλλάξουν προορισμό με επακόλουθο την πρόκληση μεγάλων καθυστερήσεων και χρηματικού κόστους [9].

Σχήμα 5: *Από-παγοποίηση αεροσκάφους, Wikipedia*

- **Ορατότητα**

Η μειωμένη ορατότητα και η χαμηλή νέφωση είναι κίνδυνοι ως προς την ασφάλεια μιας πτήσης. Σύμφωνα με την NTSB, η μειωμένη ορατότητα ήταν κύριος παράγοντας στο 24% αεροπορικών δυστυχημάτων στην περίοδο 1989-1997. Οι πύργοι ελέγχου αναγκάζονται να κατευθύνουν πολύ πιο προσεκτικά τα αεροσκάφη τόσο στον αέρα όσο και στο έδαφος με

αποτέλεσμα τις πιο αργές μετακινήσεις. Πολλές είναι οι πτήσεις που θα καθυστερήσουν να αναχωρήσουν λόγω χαμηλής ορατότητας εν αναμονή πιο καθαρού ουρανού. [9] Για τις ΗΠΑ οι κανονισμοί πτήσης για τα επιτρεπτά όρια ορατότητας καθορίζονται από τα Federal Aviation Regulations στην παράγραφο §91.155 .

Σχήμα 6: *Χαμηλή ορατότητα, Skybrary*

- **Βροχόπτωση – Χιονόπτωση**

Η έντονη βροχόπτωση ή χιονόπτωση επηρεάζει άμεσα τις καθυστερήσεις των πτήσεων, αφού απαιτείται ιδιαίτερη προσοχή στους χώρους του αεροδρομίου από τους χειριστές των αεροσκαφών και του πύργου ελέγχου που οδηγούν σε καθυστερήσεις [10].

- **Άνεμος**

Η ταχύτητα του ανέμου παρόλο που δεν επηρεάζει ιδιαίτερα μία πτήση κατά τη διάρκεια της, επηρεάζει τη διαδικασία προσγείωσης και απογείωσης. Άνεμοι ταχύτητας 15-20 m/s με οριζόντια διεύθυνση κρίνονται απαγορευτικοί για την απογείωση [10].

- **Όγκος Πτήσεων**

Ο μεγάλος όγκος των πτήσεων μπορεί να προκαλέσει συμφόρηση στο NAS και να οδηγήσει, όπως είναι λογικό, σε καθυστερήσεις. Τέτοια φαινόμενα παρατηρούνται ιδιαίτερα συγκεκριμένες μέρες και ειδικότερα τις κοντινές μέρες σε γιορτές και αργίες. Αντιθέτως, είναι λιγότερο πιθανό μια πτήση να αντιμετωπίσει καθυστέρηση μία μέρα όπου το Εθνικό Δίκτυο Αερομεταφορών έχει ελαττωμένη κίνηση [10].

1.5 Δομή εργασίας

Στο Κεφάλαιο 2 διατυπώνεται το θεωρητικό υπόβαθρο που απαιτείται για την πλήρη κατανόηση της παρούσας εργασίας.

Στο Κεφάλαιο 3 περιγράφεται η συλλογή και αξιοποίηση του συνόλου δεδομένων που χρησιμοποιείται στο πειραματικό μέρος της εργασίας.

Στο Κεφάλαιο 4 αναφέρονται οι μέθοδοι που χρησιμοποιήθηκαν για την προετοιμασία και την αξιολόγηση των πειραμάτων.

Στο Κεφάλαιο 5 δίνονται τα αποτελέσματα και ο σχολιασμός των πειραμάτων που έγιναν στο πλαίσιο της διπλωματικής εργασίας.

Στο Κεφάλαιο 6 καταγράφονται τα συμπεράσματα που εξήχθησαν από την παρούσα εργασία καθώς και κάποιες ιδέες για την μελλοντική ερευνητική επέκταση της.

Κεφάλαιο 2

ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

2.1 Μηχανική Μάθηση

Η Μηχανική Μάθηση (Machine Learning) αποτελεί έναν κλάδο της Τεχνητής Νοημοσύνης, ο οποίος αφορά την μελέτη και εφαρμογή αλγορίθμων και μοντέλων της στατιστικής για την εκτέλεση ενός έργου χωρίς προκαθορισμένες οδηγίες, αλλά με την αναγνώριση και αξιοποίηση μοτίβων και τεκμηρίων [11] .

Με τη Μηχανική Μάθηση προγραμματίζουμε υπολογιστές για τη βελτιστοποίηση ενός κριτηρίου χρησιμοποιώντας δείγμα δεδομένων ή εμπειρία από το παρελθόν [12]. Οι αλγόριθμοι οι οποίοι χρησιμοποιούνται και εκπληρώνουν τους σκοπούς της μηχανικής μάθησης, βελτιώνουν την απόδοση τους (ευστοχία πρόβλεψης – περιγραφής αντικειμένου) στην εργασία που τους έχει ανατεθεί χρησιμοποιώντας και χτίζοντας πάνω στην προηγούμενη εμπειρία που έχουν αποκτήσει. Ως εκ τούτου, οι αλγόριθμοι αυτοί έχουν την ιδιότητα να εκτελούν το έργο τους χωρίς να χρειάζεται να επαναπρογραμματιστούν κάθε φορά που παρουσιάζονται ή συλλέγονται καινούρια δεδομένα.

Οι αλγόριθμοι Μηχανικής Μάθησης αποδείχτηκαν εξαιρετικής χρησιμότητας και σημασίας σε μια πληθώρα εφαρμογών. Χρησιμοποιούνται καθημερινά σε εφαρμογές με ιατρικά δεδομένα για υποστήριξη διαγνώσεων, αλλά και για την αναγνώριση εικόνων (όπως εφαρμογές σε αυτο-οδηγούμενα οχήματα), χρηματιστηριακές προβλέψεις, συστήματα συστάσεων κ.λπ..

2.2 Κατηγορίες Μηχανικής Μάθησης

Σύμφωνα με τον τρόπο που το σύστημα επεξεργάζεται τα δεδομένα εισόδου και την ανατροφοδότηση του μοντέλου μάθησης κατά την εκπαίδευση, οι αλγόριθμοι Μηχανικής Μάθησης διακρίνονται σε τρεις ξεχωριστές κατηγορίες: [13]

1. **Επιβλεπόμενη Μάθηση (Supervised Learning):**

   Στην Επιβλεπόμενη Μηχανική Μάθηση, στόχος είναι η προσέγγιση μιας συνάρτησης η οποία συνδέει την είσοδο με την έξοδο, χρησιμοποιώντας προκαθορισμένα ζεύγη εισόδου – εξόδου (δεδομένα – ετικέτες) [13]. Τα ζεύγη αυτά έχουν καθοριστεί και δοθεί από τον χρήστη. Η διαδικασία προσέγγισης της συνάρτησης αυτής ονομάζεται εκπαίδευση. Μετά την εκπαίδευση, τελικός στόχος είναι ο υπολογισμός μιας τέτοιας συνάρτησης που να γενικεύει τα δεδομένα επαρκώς ώστε να μπορεί να αντιστοιχίσει σε ορθές εξόδους δεδομένα εισόδου τα οποία δεν χρησιμοποιήθηκαν προηγουμένως στην εκπαίδευση.

2. **Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning):**

   Σε αυτή την κατηγορία μάθησης, δεν χρησιμοποιούνται ζεύγη τιμών εισόδου-εξόδου. Στόχος είναι η αναγνώριση των μοτίβων στα δεδομένα εισόδου χωρίς την ανατροφοδότηση από τα δεδομένα εξόδου, έτσι το μοντέλο καλείται να ανακαλύψει μοτίβα και να καταλάβει από μόνο του τη δομή των χαρακτηριστικών των δεδομένων εισόδου [13]. Παράδειγμα αυτής της κατηγορίας είναι η τεχνική της Συσταδοποίησης (Clustering). Η τεχνική αυτή δημιουργεί συστάδες (clusters) δεδομένων εισόδου των οποίων τα χαρακτηριστικά είναι παρόμοια και διαφέρουν από αυτά άλλων συστάδων.

3. **Ενισχυτική Μάθηση (Reinforcement Learning):**

   Οι αλγόριθμοι αυτής της κατηγορίας αλληλοεπιδρούν με το περιβάλλον και λαμβάνουν “επιβραβεύσεις” ή “τιμωρίες” κατά τη διάρκεια της εκπαίδευσης. Αυτό οδηγεί τον αλγόριθμο προς την επιθυμητή κατεύθυνση και έτσι επιτυγχάνεται η μάθηση [13]. Χαρακτηριστικό παράδειγμα εφαρμογής μιας μάθησης αυτής της κατηγορίας είναι η αυτόνομη οδήγηση.

Στα πλαίσια της παρούσας διπλωματικής εργασίας, θα ασχοληθούμε μόνο με προβλήματα Επιβλεπόμενης Μάθησης.

2.3 Αλγόριθμοι Μηχανικής Μάθησης

Ακολουθεί παρουσίαση και επεξήγηση των διαφόρων Αλγορίθμων Μηχανικής Μάθησης που χρησιμοποιήθηκαν για τους σκοπούς της εργασίας.

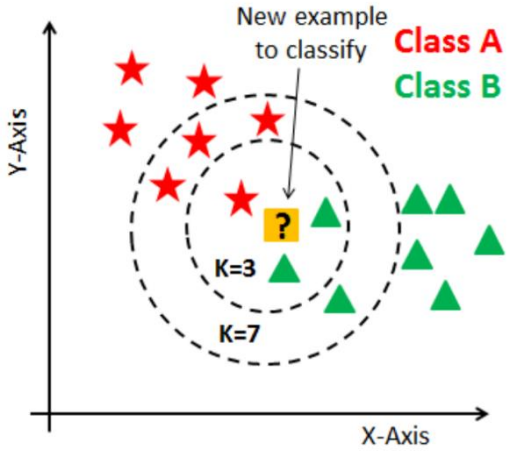
2.3.1 Κ -Κοντινότεροι Γείτονες (K-Nearest Neighbors)

Ο αλγόριθμος K-Nearest Neighbors (KNN) είναι ένας από τους πιο απλούς αλγορίθμους Μηχανικής Μάθησης. Προτάθηκε για πρώτη φορά το 1951 από τους Evelyn Fix και Joseph Hodges [14]. Ο αλγόριθμος μπορεί να χρησιμοποιηθεί για ταξινόμηση (classification) και παλινδρόμηση (regression), αλλά χρησιμοποιείται πιο συχνά σε προβλήματα ταξινόμησης. Ο αλγόριθμος αυτός βασίζεται στην αναπαράσταση των δεδομένων εισόδου ως σημεία σε ένα n-διάστατο Ευκλείδειο χώρο. Κάθε νέο δείγμα τοποθετείται στο χώρο και η τιμή εξόδου του υπολογίζεται με βάση των χαρακτηρισμό των K κοντινότερων του γειτόνων. Ο KNN είναι μη-παραμετρικός αλγόριθμος. Αυτό σημαίνει ότι δεν κάνει καμία υπόθεση για τα δεδομένα που του παρέχονται. Είναι αλγόριθμος instance-based, δηλαδή δεν πραγματοποιεί κάποια εκπαίδευση, παρά μόνο όταν κληθεί να πάρει μια απόφαση για κάποιο καινούριο δείγμα, εξετάζει εκείνη την ώρα τη σχέση του με τα ήδη αποθηκευμένα δείγματα.

Οι K κοντινότεροι γείτονες υπολογίζονται με βάση την Ευκλείδεια απόσταση των δειγμάτων (σημείων στο n-διάστατο χώρο) . Η Ευκλείδεια απόσταση που αντιστοιχεί σε δύο διανύσματα $\mathbf{x} = \{x_1, \ldots, x_n\}, \mathbf{y} = \{y_1, \ldots, y_n\}$ δίνεται ως γνωστόν από τη σχέση:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2 + \cdots + (y_n - x_n)^2} = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}.$$

Το αποτέλεσμα του αλγορίθμου είναι άμεσα συνδεδεμένο με την επιλογή του K , δηλαδή το πλήθος των γειτόνων που θα χρησιμοποιηθούν για την κατηγοριοποίηση. Η εύρεση του κατάλληλου K μπορεί να αποβεί χρονοβόρα σαν διαδικασία. Ωστόσο, γενικά, ο αλγόριθμος είναι σχετικά γρήγορος και αρκετά αποδοτικός όταν τα δεδομένα μας είναι θορυβώδη.



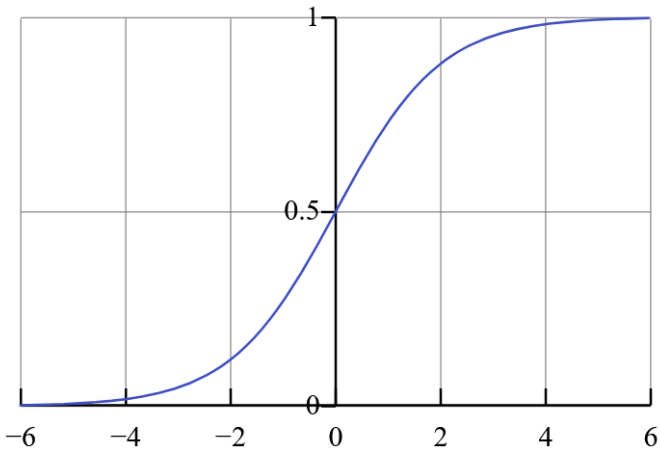
Σχήμα 7: *Ο αλγόριθμος KNN, Sarang Anil Gotke, kdnuggets*

2.3.2 Λογιστική Παλινδρόμηση (Logistic Regression)

Η Λογιστική Παλινδρόμηση είναι ένα στατιστικό μοντέλο που χρησιμεύει στην περιγραφή της σχέσης που αναπτύσσουν μία ή περισσότερες ανεξάρτητες μεταβλητές, είτε ποσοτικές συνεχείς, είτε κατηγορικές, με μια εξαρτημένη κατηγορική μεταβλητή, εκφρασμένη ως πιθανότητα δυνάμενη να πάρει καθορισμένες τιμές [14],[6]. Ανάλογα με τη φύση της εξαρτημένης κατηγορικής μεταβλητής διακρίνονται τρεις τύποι λογιστικής παλινδρόμησης, η Binary (δυαδική), η Multinomial (πολυωνυμική) και η Ordinal (τακτική). Στην παρούσα εργασία ασχολούμαστε μόνο με Binary Logistic Regression, δηλαδή η εξαρτημένη κατηγορική μεταβλητή μπορεί να πάρει δύο μόνο τιμές (π.χ. 0-1, Pass/Fail).

Το μοντέλο παίρνει το όνομα της από τη χρήση της λογιστικής συνάρτησης η οποία είναι μια σιγμοειδής συνάρτηση η οποία ορίζεται από τη σχέση:

$$S(t) = \frac{1}{1 + e^{-t}}$$

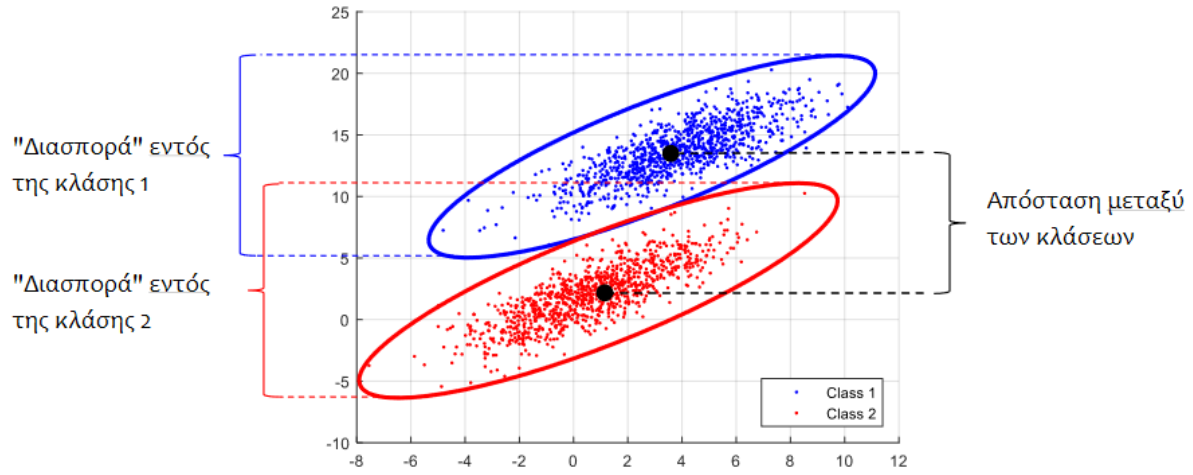


Σχήμα 8: *Σιγμοειδής Συνάρτηση,* [*Wikipedia*]

Η μέθοδος χρησιμοποιεί την Εκτίμηση Μέγιστης Πιθανότητας (MLE) αντί για τα Συνήθη Ελάχιστα Τετράγωνα (OLS) για την εκτίμηση των παραμέτρων. Επομένως, για τη σωστή εφαρμογή της Λογιστικής Παλινδρόμησης συνήθως απαιτείται μεγάλο δείγμα για την παραγωγή αξιόπιστου αποτελέσματος [15]. Λεπτομερής περιγραφή των μεθόδων της λογιστικής παλινδρόμησης παρέχεται από τα συγγράμματα των Cox & Snell (1989) [14] και των Hosmer & Lemeshow (2000) [16] .

2.3.3 Γραμμική Διακριτική Ανάλυση (Linear Discriminant Analysis)

Η Ανάλυση Γραμμικής Διάκρισης (LDA) είναι μια μέθοδος μετασχηματισμού δεδομένων τα οποία ανήκουν σε διαφορετικές κλάσεις. Η μέθοδος αυτή επιδιώκει τον καλύτερο διαχωρισμό των κλάσεων και ταυτόχρονα την ελάττωση της διαστατικότητας των δεδομένων. Στη μέθοδο LDA τα δεδομένα μετασχηματίζονται με τρόπο ο οποίος οδηγεί στην μεγιστοποίηση της απόστασης μεταξύ των κλάσεων (δηλαδή απομάκρυνση των κλάσεων) ενώ ταυτόχρονα επιδιώκεται ελαχιστοποίηση της διασποράς εντός των κλάσεων. Τα δεδομένα της εκάστοτε κλάσης δηλαδή, θα πρέπει να συγκεντρώνονται γύρω από τη μέση τιμή τους. Βασική προϋπόθεση για την εφαρμογή LDA είναι πως τα δεδομένα πρέπει να ορίζονται με τέτοιο τρόπο ώστε να χωρίζονται σε δύο ή περισσότερες αμοιβαία αποκλειόμενες ομάδες, δηλαδή κάθε δείγμα ανήκει σε μοναδική ομάδα.



Σχήμα 9: *Απεικόνιση Κλάσεων, I. Kalatzis*

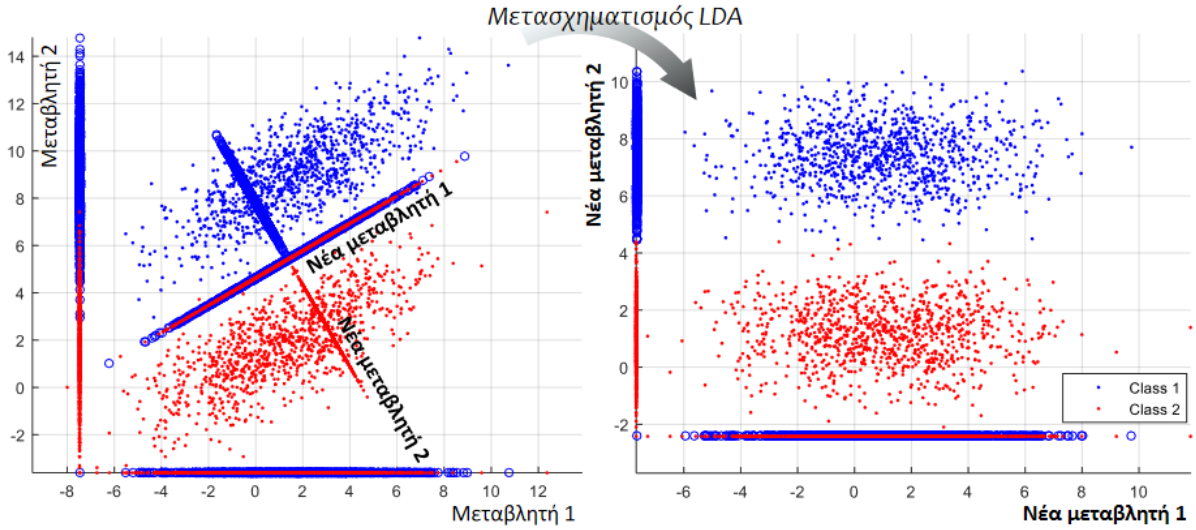
Κατά το μετασχηματισμό γίνεται προβολή των δεδομένων σε χώρο μικρότερης διαστατικότητας από τον αρχικό, οδηγώντας έτσι σε ελάττωση στο πλήθος των διαστάσεων τους, δηλαδή το πλήθος των μεταβλητών από τις οποίες αποτελούνται τα δεδομένα. Με αυτό τον τρόπο επιτυγχάνεται η καλύτερη διαχωρισιμότητα τους σε κλάσεις, έτσι ώστε στη συνέχεια να έχουμε ακριβέστερη και ευκολότερη ταξινόμηση νέων δεδομένων στις κλάσεις. Για τον καλό διαχωρισμό κλάσεων χρησιμοποιείται το κριτήριο Fisher όπως αυτό αποδίδεται από τον τύπο [17] :

$$J = \frac{διασπορά\ μεταξύ\ των\ κλάσεων}{διασπορά\ εντός\ των\ κλάσεων} = \frac{S_{between}}{S_{within}}$$

Όταν το κριτήριο Fisher μεγιστοποιείται, τότε επιτυγχάνεται καλύτερος διαχωρισμός, που κατ' επέκταση σημαίνει:

- Μεγιστοποίηση της απόστασης μέσων τιμών των κλάσεων, και ταυτόχρονα
- Ελαχιστοποίηση της διασποράς εντός κάθε κλάσης (συγκέντρωση δεδομένων κλάσης γύρω από τη μέση τιμή της)

Με την κατάλληλη προβολή ο λόγος αυτός μεγιστοποιείται και ο διαχωρισμός γίνεται βέλτιστος.



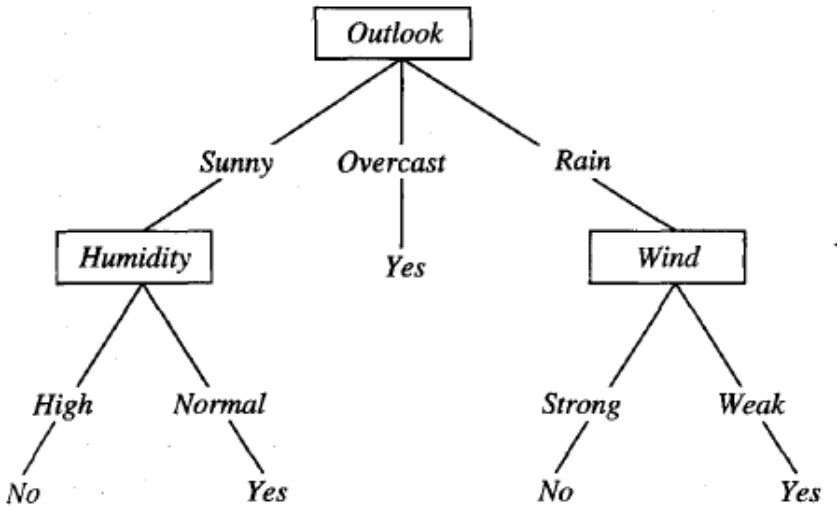
Σχήμα 10: *Μετασχηματισμός LDA, I. Kalatzis*

Μία παρόμοια μέθοδος που χρησιμοποιείται είναι η Quadratic Discriminant Analysis (QDA), η οποία παρέχει μια εναλλακτική προσέγγιση υποθέτοντας ότι κάθε κλάση έχει το δικό της πίνακα συνδιακύμανσης Σk, εν αντιθέσει με την LDA που υποθέτει κοινό πίνακα συνδιακύμανσης για όλες τις K κλάσεις.

2.3.4 Δέντρα Αποφάσεων (Decision Trees)

Τα Δέντρα Αποφάσεων είναι ευρέως διαδεδομένα και χρησιμοποιούμενα σε προβλήματα ταξινόμησης. Μπορούν να εφαρμοστούν σε μεγάλη ποικιλία προβλημάτων ταξινόμησης, όπως π.χ. για τη λήψη απόφασης με βάση πιθανές εκβάσεις σε ένα παίγνιο, την ιατρική διάγνωση βασισμένη σε συμπτώματα, τη διαπίστωση πιθανότητας καταιγίδας σύμφωνα με καιρικές παρατηρήσεις, κ.ά. Ο αλγόριθμος αυτός, οδηγεί στη δημιουργία μιας ανάποδης δενδροειδούς δομής στην οποία οι κατηγορίες ταξινόμησης είναι τα φύλλα. Τα

δέντρα αυτά μπορούν να αναπαρασταθούν ως σύνολο κανόνων if-then για να είναι πιο αναγνώσιμα από τον άνθρωπο. Τα Δέντρα Αποφάσεων ταξινομούν ένα δείγμα πραγματοποιώντας μια διάσχιση του δέντρου με αφετηρία τη ρίζα και τερματισμό σε κάποιο από τα φύλλα. Η διάσχιση του δέντρου γίνεται "απαντώντας" κατάλληλα σε κάθε κόμβο του δέντρου και συνεχίζοντας στο σωστό μονοπάτι. Κάθε κόμβος του δέντρου αποτελεί έλεγχο ενός χαρακτηριστικού του δείγματος και κάθε κλαδί αυτού του κόμβου αντιστοιχεί σε μια από τις πιθανές τιμές αυτού του χαρακτηριστικού [18],[19]. Για την καλύτερη κατανόηση, παρατίθεται το παρακάτω παράδειγμα δέντρου απόφασης το οποίο σκοπεύει στην λήψη απόφασης για την πραγματοποίηση ενός αγώνα τένις, λαμβάνοντας υπόψιν τις καιρικές συνθήκες.



Σχήμα 11: *Decision Tree PlayTennis, Machine Learning, Tom Mitchell*

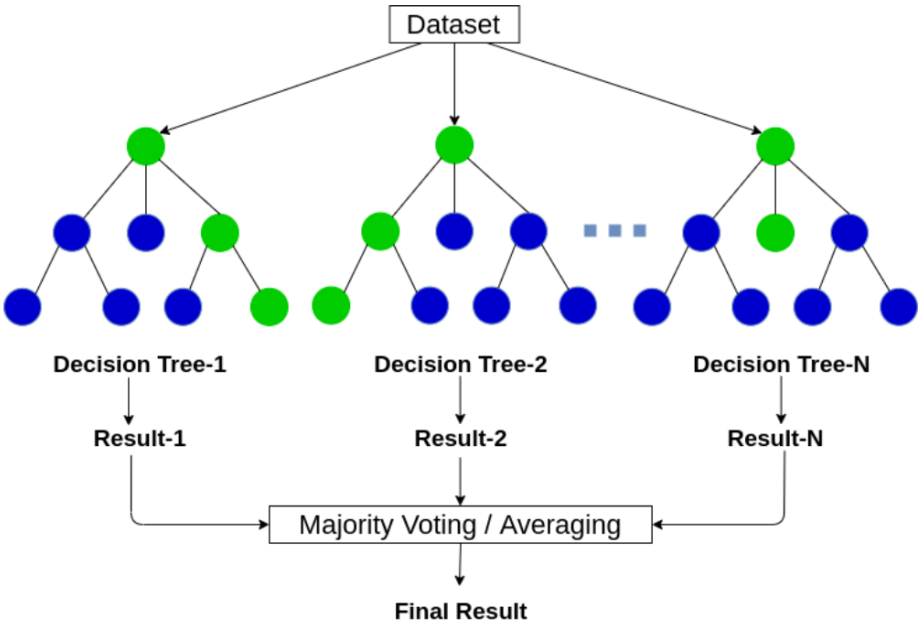
Σε ένα τέτοιο αλγόριθμο μπορούμε εύκολα να φτάσουμε σε υπερπροσαρμογή (overfitting), αφού υπάρχει η πιθανότητα δημιουργίας μεγάλου πλήθους κλαδιών λόγω θορύβου ή ακραίων τιμών (outliers). Σε μια τέτοια περίπτωση παρόλο που το μοντέλο μας θα έχει πολύ καλή απόδοση στα δεδομένα εκπαίδευσης θα έχει κακή απόδοση σε καινούρια δεδομένα.

2.3.5 Τυχαία Δάση (Random Forests)

Ένας από τους πιο αποδοτικούς αλγόριθμους ταξινόμησης ML είναι ο αλγόριθμος των Τυχαίων Δασών (*Random Forests*). Προτάθηκε για πρώτη φορά από τον Tin Kam Ho [21],

βελτιώθηκε αργότερα από τον Breiman [22] και είναι πλέον από τους πιο γνωστούς και διαδεδομένους αλγόριθμους. Ο αλγόριθμος RF ακολουθεί μια συνηθισμένη τεχνική στην Επιβλεπόμενη Μάθηση, την ανεξάρτητη εκπαίδευση πολλών ταξινομητών και το συνδυασμό των αποτελεσμάτων τους για την τελική πρόβλεψη (Ensemble Learning) [23]. Η χρήση αυτής της τεχνικής επιφέρει σημαντικά πλεονεκτήματα καθώς λόγω της ανεξάρτητης εκπαίδευσης του κάθε ταξινομητή, ο καθένας μπορεί να επικεντρωθεί σε κάποια χαρακτηριστικά των δεδομένων και να αδιαφορήσει για άλλα, τα οποία θα αξιοποιήσει κάποιος άλλος ταξινομητής. Μετά την ξεχωριστή εκπαίδευση ακολουθεί μια συνάρτηση συσχέτισης η οποία θα αξιοποιήσει τα επιμέρους αποτελέσματα και θα μας δώσει την τελική πρόβλεψη [21].

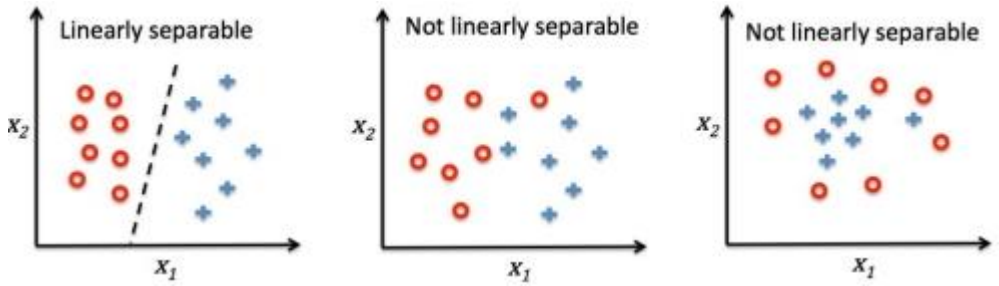
Ένα Random Forest αποτελεί μια συλλογή από πολλά Decision Trees. Κατά την διάρκεια της εκπαίδευσης των δέντρων χρησιμοποιείται μια τεχνική αποκαλούμενη "σακούλιασμα - bagging" η οποία ορίζει σε κάθε δέντρο το υποσύνολο δεδομένων εκπαίδευσης που θα χρησιμοποιήσει κατά την εκπαίδευση του. Αυτό επιτρέπει στους ταξινομητές Decision Trees να αντιλαμβάνονται καλύτερα συγκεκριμένα χαρακτηριστικά. Με το πέρας της εκπαίδευσης του, το κάθε Tree δίνει ένα αποτέλεσμα, και ως εκ τούτου, η κάθε κλάση υποδεικνύεται από ορισμένο αριθμό trees. Στη συνέχεια η τελική και οριστική ταξινόμηση γίνεται με την επιλογή της επικρατέστερης κλάσης από το Τυχαίο Δάσος [21], [22] .



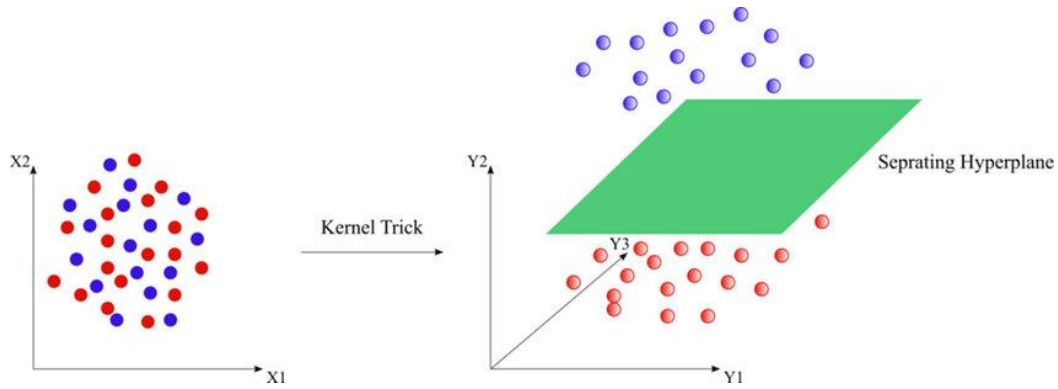
Σχήμα 12: *Random Forest, Abhishek Sharma, Analytics Vidhya*

2.3.6 Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)

Ο αλγόριθμος αυτός βασίζεται στην αναπαράσταση των δεδομένων εισόδου στο χώρο. Δημιουργώντας ένα κατάλληλο υπερεπίπεδο (hyperplane), ο αλγόριθμος διαχωρίζει τα δεδομένα στις κλάσεις εξόδου. Το υπερεπίπεδο πρέπει να διαχωρίζει τις κλάσεις με τη μεγαλύτερη δυνατή ακρίβεια, ενώ παράλληλα να μεγιστοποιεί την απόσταση των δεδομένων εισόδου κάθε κλάσης ως προς αυτό. Ανάλογα με τα δεδομένα εισόδου, οι κατηγορίες εξόδου μπορεί να είναι γραμμικά διαχωρίσιμες ή μη-γραμμικά διαχωρίσιμες. Στην δεύτερη περίπτωση στην οποία τα δεδομένα εισόδου δεν είναι γραμμικά διαχωρίσιμα, εφαρμόζονται κατάλληλοι μετασχηματισμοί στα δεδομένα με χρήση συγκεκριμένων συναρτήσεων πυρήνα (kernel functions) με στόχο τη μεταφορά τους σε ένα νέο χώρο άλλων διαστάσεων. Σημαντικό χαρακτηριστικό των SVM είναι ότι δεν έχουν πρόβλημα στην κατηγοριοποίηση δεδομένων πολλών διαστάσεων.



Σχήμα 13: *Παράδειγμα Γραμμικά Διαχωρίσεων και Μη Γραμμικά Διαχωρίσιμων Κλάσεων, Sebastian Rashka,*



Σχήμα 14*: SVM Classification for Non-Linearly Separable Data-Points, Arash Saeidpour*

Όπως αναφέρθηκε τα δεδομένα εισόδου πρέπει να διαχωριστούν από κάποιο βέλτιστο υπερεπίπεδο. Τα κοντινότερα σημεία σε αυτό το υπερεπίπεδο ονομάζονται Διανύσματα Υποστήριξης (Support Vectors) και φέρουν τη μεγαλύτερη πρόκληση ως προς την ταξινόμηση τους. Με κριτήριο την μεγιστοποίηση την απόσταση τους από το υπερεπίπεδο, επηρεάζουν άμεσα την βέλτιστη τοποθεσία του υπερεπιπέδου.

Το αποτέλεσμα ενός SVM καθορίζεται άμεσα από την επιλογή της συνάρτησης πυρήνα. Ενδεικτικά κάποιες από αυτές τις συναρτήσεις τις οποίες θα συναντήσουμε και στο πειραματικό μέρος είναι:

I. Polynomial

$$k(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i} \cdot \mathbf{x_j} + 1)^d$$

II. Gaussian Radial basis function (RBF)

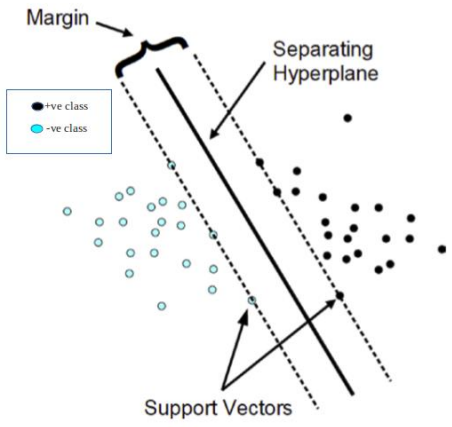
$$k(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\gamma \|\mathbf{x_i} - \mathbf{x_j}\|^2)_{\text{για}} \quad \gamma > 0$$

III. Hyperbolic Tangent

$$k(\mathbf{x_i}, \mathbf{x_j}) = \tanh(\kappa \mathbf{x_i} \cdot \mathbf{x_j} + c)$$

IV. Sigmoid

$$k(x, y) = \tanh(\alpha x^T y + c)$$



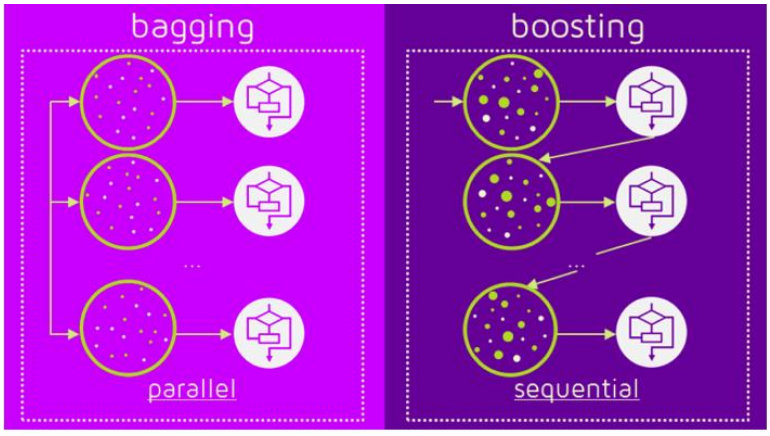
Σχήμα 15: *Support Vector Machine, Yash Rawat*

2.4 Ενίσχυση

2.4.1 Η Τεχνική της Ενίσχυσης (Boosting)

Παραδοσιακά, στην Μηχανική Μάθηση ένα πρόβλημα μπορούσε να αντιμετωπιστεί με μόνο ένα μοντέλο (single learner) όπως αυτά που είδαμε στο προηγούμενο υποκεφάλαιο (Decision Trees, Support Vector Machines κ.λπ.). Στην πορεία όμως στην αναζήτηση για καλύτερα αποτελέσματα, γεννήθηκαν τεχνικές που χρησιμοποιούν σύνολο μοντέλων (**Ensemble Methods**) [23]. Οι μέθοδοι αυτοί χρησιμοποιούν πολλούς learners για να αυξήσουν την απόδοση των μοντέλων. Αυτές οι μέθοδοι μπορούν να χαρακτηριστούν ως τεχνικές που βασίζονται σε ένα σύνολο από **αδύναμους (weak) learners**, δηλαδή learners που επιτυγχάνουν οριακά καλύτερα αποτελέσματα από τυχαία ταξινόμηση, με σκοπό τη δημιουργία ενός πιο **δυνατού (strong) learner**.

Η Ενίσχυση (Boosting) είναι μια μέθοδος Ensemble με κυριότερο σκοπό τη μείωση της μεροληψίας (bias) και της διακύμανσης (variance). Στην επιβλεπόμενη μάθηση, αυτό μεταφράζεται σε αλγόριθμους που μετατρέπουν weak learners σε strong learners. Η Ενίσχυση ως τεχνική βασίζεται σε ερώτημα που τέθηκε από τους Kearns και Valiant (1988, 1989), [23] “Μπορεί ένα σύνολο από weak learners να δημιουργήσουν ένα μοναδικό stronger learner ;”. Το ερώτημα απαντήθηκε καταφατικά από τον Robert Schapire το 1990 [24].



Σχήμα 16: *Difference between Bagging and Boosting, Sai Nikhilesh Kasturi*

Είναι σημαντικό να γίνει σαφής η διάκριση της τεχνικής Ενίσχυσης από την τεχνική Σακουλιάσματος (Bagging), που είναι η άλλη κύρια οικογένεια των μεθόδων Ensemble και την οποία έχουμε συναντήσει προηγουμένως στα Random Forests (RF). Ενώ στο Bagging οι

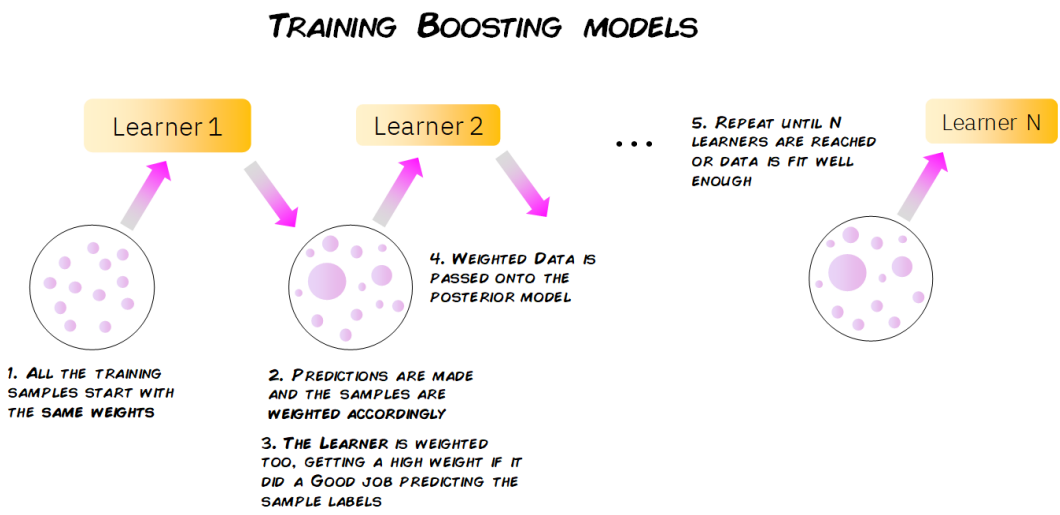
learners εκπαιδεύονται παράλληλα, στο Boosting οι learners εκπαιδεύονται διαδοχικά για να πετύχουν τον σκοπό που αναφέρθηκε προηγουμένως.

2.4.2 Αλγόριθμοι Boosting

Οι περισσότεροι αλγόριθμοι Ενίσχυσης αποτελούνται από αδύναμους learners που εκπαιδεύονται επαναληπτικά με βάση κάποια κατανομή και συγκεντρώνονται σε έναν τελικό δυνατό learner. Αφού συγκεντρωθούν, σταθμίζονται με τρόπο ανάλογο της ακρίβειας των αδύναμων learners. Με την προσθήκη ενός αδύναμου learner, τα βάρη των δεδομένων επαναπροσδιορίζονται, μια διαδικασία γνωστή ως “re-weighting”. Δεδομένα που έχουν ταξινομηθεί λανθασμένα αποκτούν μεγαλύτερο βάρος και δεδομένα με σωστή ταξινόμηση, χάνουν από το βάρος τους. Ως εκ τούτου, μελλοντικοί αδύναμοι learners εστιάζουν περισσότερο σε δείγματα τα οποία προηγούμενοι αδύναμοι learners ταξινόμησαν λανθασμένα. Υπάρχουν αρκετοί αλγόριθμοι Ενίσχυσης. Τους πρώτους πρότειναν οι Robert Schapire και Yoav Freund, αλλά δεν ήταν προσαρμοστικοί και δεν μπορούσαν να εκμεταλλευτούν πλήρως τους αδύναμους learners. Στη συνέχεια όμως, ο Schapire και ο Freund κατασκεύασαν τον αλγόριθμο **Προσαρμοσμένης ενίσχυσης – AdaBoost**, για τον οποίον βραβεύτηκαν με το υψηλού κύρους βραβείο Gödel το 2003 [25], [26].

Σήμερα υπάρχουν τεχνικές Ενίσχυσης που είναι πολύ δημοφιλείς και χρησιμοποιούνται σε πληθώρα εφαρμογών. Κάποιοι από αυτούς θα μας απασχολήσουν στο πειραματικό κομμάτι της εργασίας:

- Προσαρμοσμένη ενίσχυσης – AdaBoost
- Ενίσχυση Κλίσης – Gradient Boosting
- Ακραία Ενίσχυση Κλίσης – Extreme Gradient Boosting

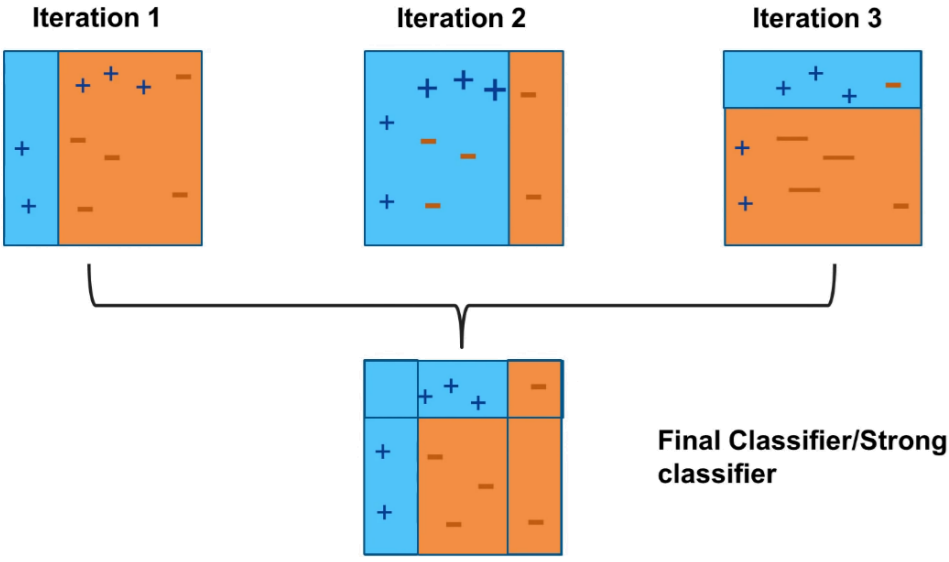


Σχήμα 17: *Training Boosting Models, z_ai*

2.4.3 AdaBoost

Ο AdaBoost, ο αλγόριθμος Προσαρμοσμένης Ενίσχυσης, είναι ένας αλγόριθμος στατιστικής που χρησιμοποιείται για ταξινόμηση [26]. Μπορεί να χρησιμοποιηθεί σε συνδυασμό με διάφορους αλγόριθμους ως βάση του για να βελτιωθεί η απόδοση του μοντέλου. Το αποτέλεσμα των άλλων αλγόριθμων μάθησης (weak learners) συνδυάζεται σε ένα σταθμισμένο άθροισμα που αντιπροσωπεύει την τελική έξοδο του ενισχυμένου ταξινομητή. Ο AdaBoost είναι προσαρμοστικός υπό την έννοια ότι κάθε weak learner που διαδέχεται έναν άλλο, τροποποιείται υπέρ των δειγμάτων που ταξινομήθηκαν λανθασμένα από προηγούμενους learners. Η απόδοση αυτών των learners μπορεί να είναι πολύ κακή, αρκεί όμως να είναι καλύτερη από τυχαία ταξινόμηση για να οδηγηθεί το μοντέλο σε σύγκλιση, σε έναν δυνατό learner. Η πιο συνηθισμένη υλοποίηση του AdaBoost είναι ο συνδυασμός του με Decision Trees ως weak learners και θεωρείται από τους κορυφαίους έτοιμους (“out of the box”) αλγόριθμους ταξινόμησης.

Σε αυτή την υλοποίηση έχουμε την χρήση Decision Tree Stumps, τα οποία είναι ουσιαστικά Δέντρα Απόφασης που αποτελούνται από τη ρίζα και δύο φύλλα. Στο κάθε Stump αξιολογείται μόνο ένα χαρακτηριστικό των δεδομένων εισόδου. Όπως μπορούμε να συμπεράνουμε, λαμβάνοντας υπόψιν μόνο ένα χαρακτηριστικό, κάθε stump είναι πολύ αδύναμος learner. Με το συνδυασμό όμως μεγάλου αριθμού αυτών, επιτυγχάνεται η κατασκευή ενός αξιόπιστου και μεγάλης ακρίβειας Ensemble μοντέλου ταξινόμησης.



Σχήμα 18: *AdaBoost, Packt Video*

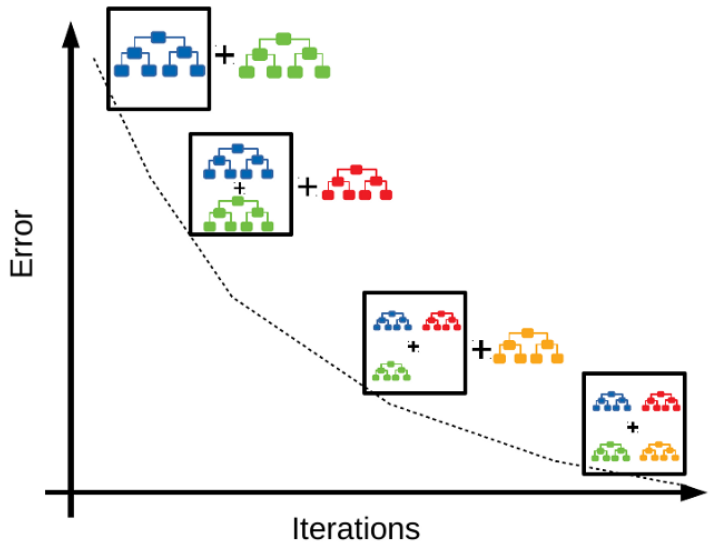
2.4.4 Ενίσχυση Κλίσης και XGBoost

Ο αλγόριθμος Ακραίας Ενίσχυσης Κλίσης (Extreme Gradient Boosting), βασίζεται στα δέντρα απόφασης και χρησιμοποιεί την τεχνική ενίσχυσης κλίσης για την βελτίωση της επίδοσης του μοντέλου. Ο αλγόριθμος αυτός παρουσιάζει αρκετές ομοιότητες με την Προσαρμοσμένη Ενίσχυση που είδαμε προηγουμένως. Γίνεται χρήση αδύναμων learners οι οποίοι συνδυάζονται με σταθμισμένα βάρη κατά την εκπαίδευση και συγκεντρώνονται με παρόμοιο τρόπο για την δημιουργία του δυνατού learner που παράγει την τελική τιμή εξόδου.

Αρχικά, δημιουργείται ένα μοντέλο βασισμένο σε ένα υποσύνολο των δεδομένων. Με αυτό το μοντέλο γίνονται προβλέψεις σε ολόκληρο το σύνολο των δεδομένων εκπαίδευσης και κατόπιν υπολογίζεται το σφάλμα. Το αδύναμο μοντέλο παράγει ανεπαρκείς προβλέψεις και έτσι πρέπει να ενισχυθεί σε μεταγενέστερες επαναλήψεις. Έτσι δημιουργείται ένα καινούριο μοντέλο το οποίο λαμβάνει υπόψιν τα σφάλματα που υπολογίστηκαν ήδη και επιχειρεί να εξαλείψει τα λάθη του προηγούμενου. Οι προβλέψεις αυτής της νέας επανάληψης συνδυάζονται με τις προβλέψεις της προηγούμενης [27].

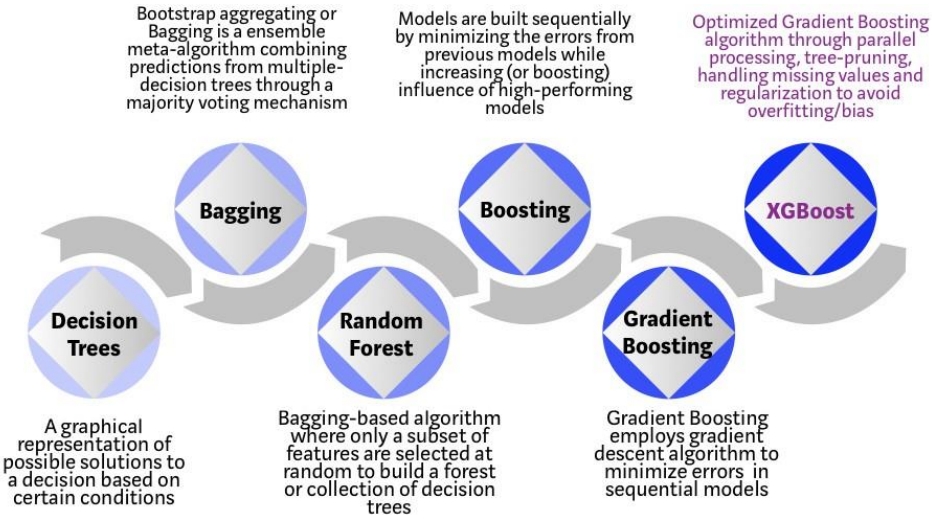
Η διαφορά της Ενίσχυσης Κλίσης από την Προσαρμοσμένη Ενίσχυση είναι ότι αυτή τη φορά, αντί να αλλάξει η βαρύτητα των στιγμιότυπων όπου ταξινομήθηκαν λάθος, γίνεται εκπαίδευση κάθε νέου μοντέλου αξιοποιώντας τα υπολειπόμενα σφάλματα του προηγούμενου.

Η επαναληπτική αυτή διαδικασία σταματά είτε όταν το σφάλμα δεν αλλάζει, ή αν επιτευχθεί το μέγιστο όριο του αριθμού των μοντέλων.

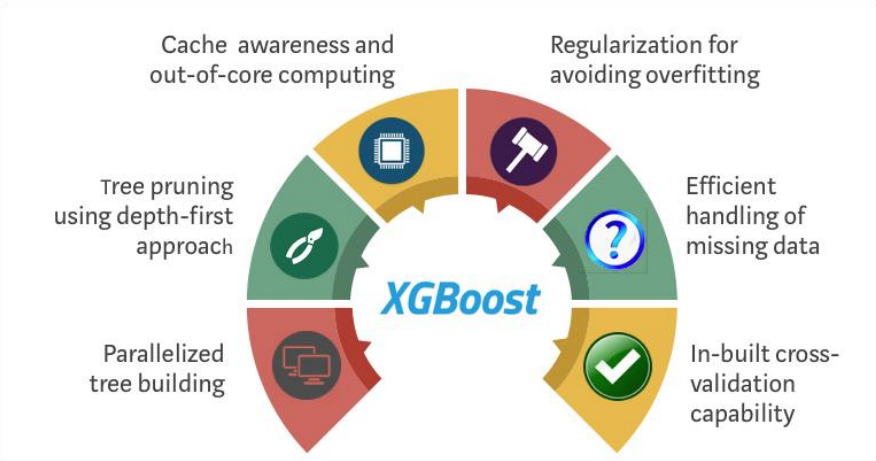


Σχήμα 19: *Παράδειγμα αλγόριθμου Ενίσχυσης Κλίσης, Aratrika Pal*

Σε προβλήματα πρόβλεψης όπου τα δεδομένα είναι αδόμητα (εικόνες, κείμενο) τα Τεχνητά Νευρωνικά Δίκτυα (ΤΝΔ) τείνουν να επικρατούν των υπόλοιπων αλγορίθμων. Ωστόσο σε δεδομένα που είναι δομημένα, οι αλγόριθμοι που βασίζονται στα δέντρα απόφασης είναι συνήθως καλύτεροι. Ο XGBoost είναι μια εξέλιξη των δέντρων απόφασης που χρησιμοποιεί αρκετές έξυπνες βελτιστοποιήσεις για να πετυχαίνει καλύτερο αποτέλεσμα. Κάποιες από τις βελτιστοποιήσεις που χρησιμοποιεί είναι παραλληλοποίηση, κλάδεμα δέντρων, και βελτιστοποίηση υπολογιστικών πόρων.



Σχήμα 20: *Εξέλιξη αλγόριθμων βασισμένων σε δέντρα αποφάσεων, XGBoost Documentation*

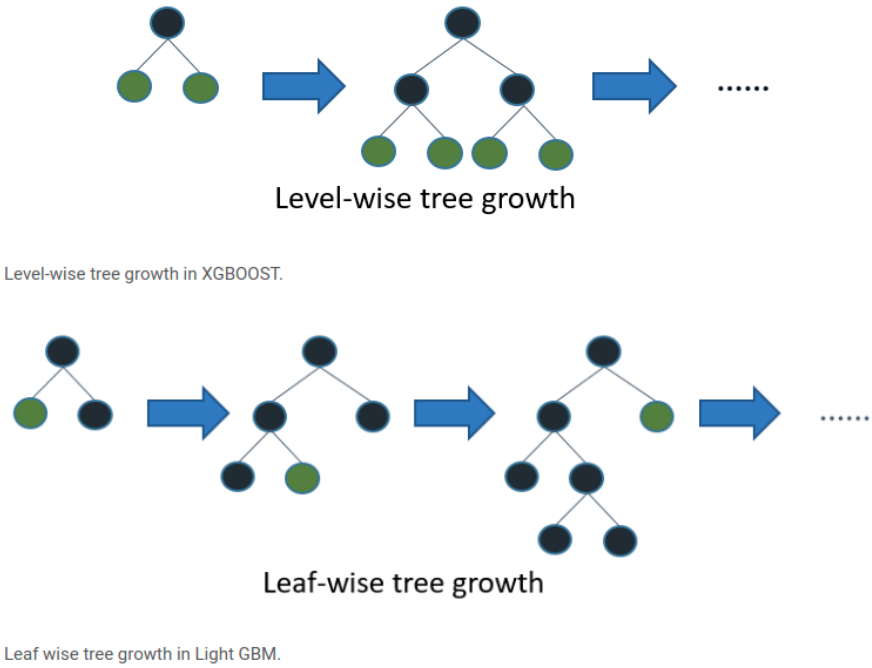


Σχήμα 21: *XGBoost, XGBoost Documentation*

Ο αλγόριθμος αυτός δημιουργήθηκε ως μέρος έρευνας στο Πανεπιστήμιο της Washington. Οι Tianqi Chen και Carlos Guestrin παρουσίασαν τη δουλειά τους στο συνέδριο SIGKDD 2016 [28].

2.4.5 LightGBM

Ο Light Gradient Boosting Machine είναι άλλος ένας αλγόριθμος που βασίζεται σε δέντρα αποφάσεων και ενίσχυση κλίσης. Ο αλγόριθμος αυτός δημιουργήθηκε από τον Guolin Ke στην Microsoft [29]. Ο LightGBM διαφέρει ως προς το γεγονός ότι αναπτύσσει δέντρα οριζόντια, δηλαδή επιλέγει να αναπτύξει το φύλλο που πιστεύει θα οδηγήσει στη μεγαλύτερη μείωση της απώλειας, ενώ άλλοι αλγόριθμοι αναπτύσσουν δέντρα ανά επίπεδο, βασιζόμενοι στο βάθος του δέντρου. Στο *Σχήμα 22*, φαίνεται μια απεικόνιση της διαφοράς του από τους υπόλοιπους αλγορίθμους της κατηγορίας του.

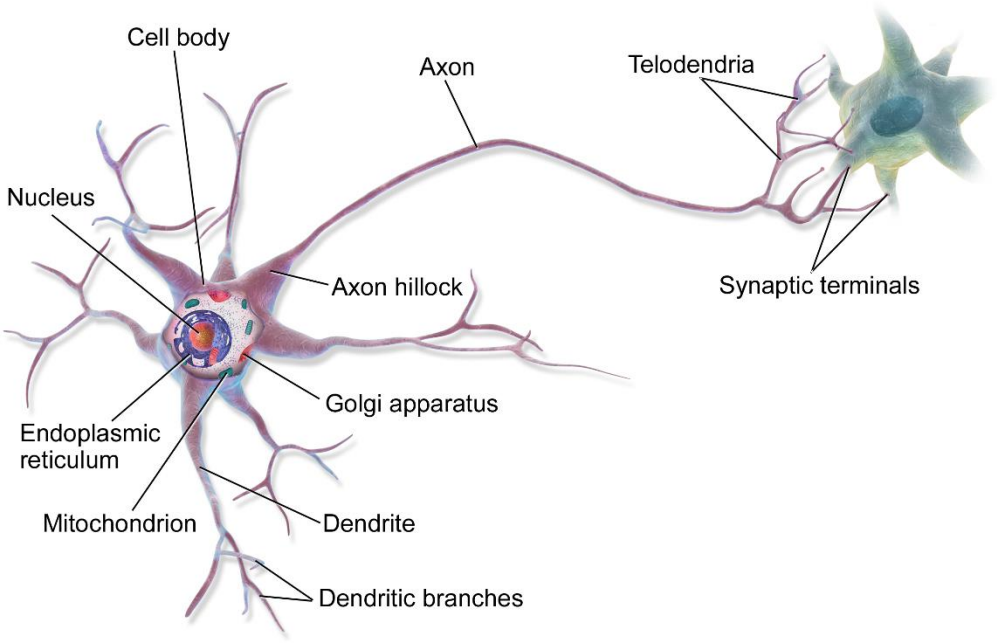


Σχήμα 22: *Διαφορά στην ανάπτυξη Δέντρου μεταξύ XGBoost και LightGBM, LightGBM Documentation*

Επιπλέον, δεν χρησιμοποιεί τον κλασσικό αλγόριθμο των δέντρων αποφάσεων, ο οποίος ψάχνει το καλύτερο σημείο διαίρεσης βασισμένος σε ταξινομημένες τιμές χαρακτηριστικών των δεδομένων, αλλά υλοποιεί έναν εξαιρετικά βελτιστοποιημένο αλγόριθμο δέντρων αποφάσεων που βασίζεται σε ιστογράμματα. Βασικά πλεονεκτήματα του αλγόριθμου είναι η μικρή χρήση μνήμης, η παράλληλη μάθηση και η ευκολία του να διαχειρίζεται μεγάλους όγκους δεδομένων.

2.5 Τεχνητά Νευρωνικά Δίκτυα

Τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks, ANN) είναι αλγόριθμοι που μιμούνται τη βιολογική δομή του εγκεφάλου. Προγραμματίζονται χωρίς συγκεκριμένους κανόνες και δεν υιοθετούν ειδικά σχεδιασμένους αλγόριθμους αναζήτησης. Τα μοντέλα αυτά "μαθαίνουν" να εκτελούν εργασίες που τους ανατίθενται εξετάζοντας παραδείγματα με σκοπό τη γενίκευση των συμπερασμάτων που θα εξάγουν απ' αυτά. Ένα ΤΝΔ βασίζεται σε μια συλλογή από τεχνητούς νευρώνες, οι οποίοι μοντελοποιούν τις ιδιότητες ενός βιολογικού νευρώνα.

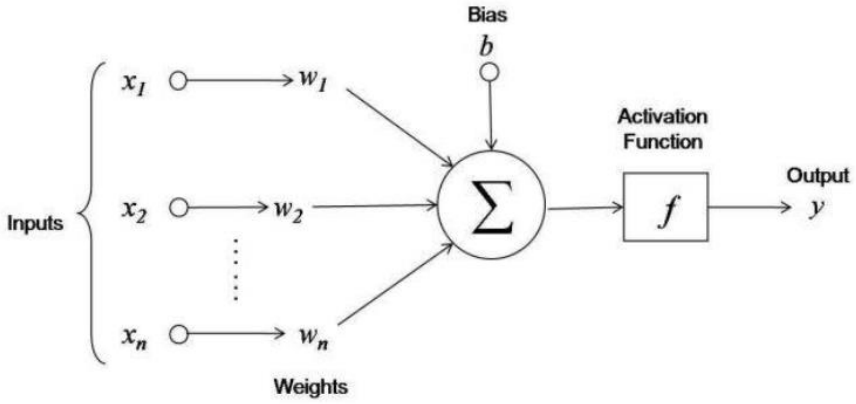


Σχήμα 23: *Βιολογικός Νευρώνας*

Ένας βιολογικός νευρώνας αποτελεί την κύρια λειτουργική μονάδα του νευρικού συστήματος και έχει τη δυνατότητα να μεταδίδει ηλεκτρικά σήματα από το ένα μέρος του κυττάρου στο άλλο. Η σηματοδότηση, γίνεται μέσω της μεταβίβασης της νευρικής ώσης (σήματος) από τους δενδρίτες. Ένα σήμα, θα πυροδοτήσει μια σειρά αντιδράσεων, οι οποίες για να πραγματωθούν, είτε θα ελέγχονται από μεμονωμένα νευρικά κύτταρα/ νευρώνες, είτε θα διαθωθούν μέσω των συνάψεων σε άλλους, γειτονικούς νευρώνες οι οποίοι θα αποκριθούν συνδυαστικά. Οι νευρώνες μαζί με τις νευρωνικές συνάψεις τους (αισθητικές ή κινητικές), σχηματίζουν νευρωνικά δίκτυα.

2.5.1 Perceptron

Σε αντιστοιχία με τον βιολογικό νευρώνα, ο τεχνητός νευρώνας αποτελείται από το σώμα το οποίο λαμβάνει τα σήματα εισόδου. Τα σήματα αυτά μετασχηματίζονται με βάση τα βάρη της εκάστοτε εισόδου και την πόλωση (bias), αθροίζονται και περνάνε από κάποια συνάρτηση ενεργοποίησης και παράγεται μοναδική έξοδος.



Σχήμα 24: *Τεχνητός Νευρώνας*

Η λειτουργία του τεχνητού νευρώνα φαίνεται στο *Σχήμα 24*. Οι είσοδοι x_i πολλαπλασιάζονται με τα αντίστοιχα βάρη w_i. Στο σώμα του νευρώνα αθροίζονται και περνάνε από μια συνάρτηση ενεργοποίησης αφού έχει προστεθεί και το bias. Η συνάρτηση ενεργοποίησης θα δώσει τιμή εξόδου 1 ή 0, δηλαδή αν ο νευρώνας ενεργοποιείται ή απενεργοποιείται. Η έξοδος προκύπτει ως:

$$y = f(\sum_{i=1}^{n} x_i w_i + b)$$

όπου f, η συνάρτηση ενεργοποίησης (βλ. 2.5.3) [30].

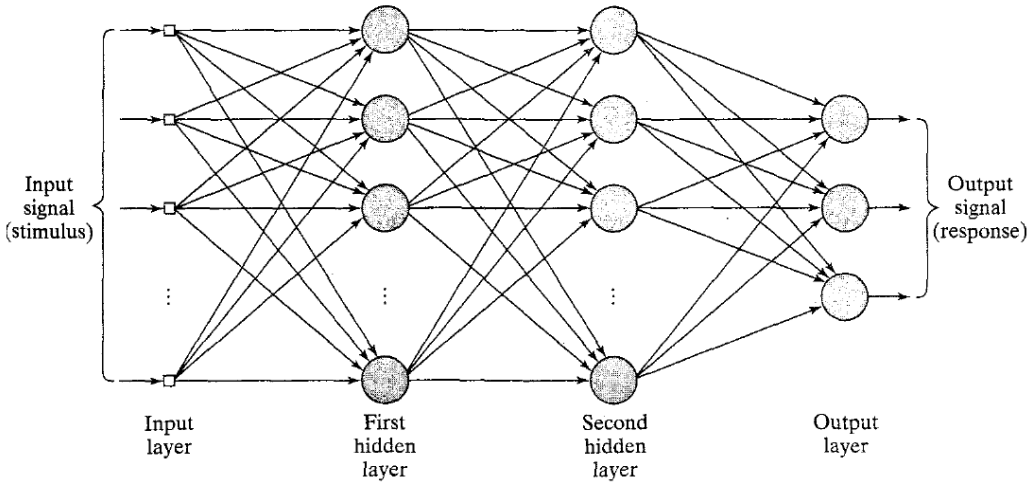
Ο τεχνητός νευρώνας είναι η βάση του αλγόριθμου Perceptron, ο οποίος προτάθηκε το 1958 από τον Frank Rosenblatt στο Cornell Aeronautical Laboratory. Αυτός ο αλγόριθμος επιβλεπόμενης μάθησης είναι ένας δυαδικός ταξινομητής ο οποίος μπορεί να αποφασίσει εάν μια είσοδος η οποία αναπαρίσταται από ένα διάνυσμα ανήκει σε μια καθορισμένη κλάση με τη μέθοδο που αναφέρθηκε παραπάνω [31].

Ο συνδυασμός περισσοτέρων από ένα Perceptron δημιουργεί ένα επίπεδο απλών νευρώνων το οποίο αποτελεί ταυτόχρονα την είσοδο και την έξοδο του δικτύου. Το 1969 στο διάσημο βιβλίο «Perceptrons» οι Minsky και Papert απέδειξαν ότι ένα δίκτυο Perceptron μονού επιπέδου (single-layer perceptron) είναι ικανό να επιλύσει μόνο γραμμικά διαχωρίσιμα προβλήματα [32]. Για πιο περίπλοκα προβλήματα είναι αναγκαία η δημιουργία Πολυεπίπεδων Νευρωνικών Δικτύων.

2.5.2 Πολυεπίπεδα Perceptron (Multi-Layer Perceptron)

Ο συνδυασμός πολλών νευρώνων μεταξύ τους για τη δημιουργία πολυεπίπεδων αρχιτεκτονικών Perceptrons αποκαλούνται **Multi-Layer Perceptrons (MLP)**. Σε ένα MLP υπάρχουν τουλάχιστον τρία επίπεδα νευρώνων, όπου οι νευρώνες ανήκουν σε ένα από τρία διαφορετικά είδη επιπέδου:

- Επίπεδο εισόδου (input layer)
- Κρυφά επίπεδα (hidden layers)
- Επίπεδο Εξόδου (output layer)



Σχήμα 25: *Architectural graph of a multilayer perceptron with two hidden layers, Simon Haykin*

Το επίπεδο εισόδου αναλαμβάνει την εισαγωγή δεδομένων στο MLP. Ακολουθούν τα κρυφά επίπεδα τα οποία συνδέουν το προηγούμενο και το επόμενο τους επίπεδο. Δεν υπάρχει περιορισμός στο πλήθος των νευρώνων σε ένα κρυφό επίπεδο καθώς και ο αριθμός των επιπέδων αυτών, είναι καθαρά σχεδιαστική επιλογή. Τέλος υπάρχει το επίπεδο εξόδου όπου

εμφανίζονται τα τελικά αποτελέσματα που προέκυψαν από τη διαδικασία εκπαίδευσης. Ο αριθμός των νευρώνων του επιπέδου εξόδου ισούται με το πλήθος των πιθανών εξόδων. Ανάλογα με τον τρόπο που συνδέονται οι νευρώνες σε ένα ΤΝΔ μπορούμε να χαρακτηρίσουμε το δίκτυο μας ως:

- **Πλήρως συνδεμένο (fully connected):** Όλοι οι νευρώνες ενός επιπέδου είναι συνδεδεμένοι με κάθε νευρώνα σε ένα διαφορετικό επίπεδο
- **Μερικώς συνδεδεμένο (partially connected):** Δεν είναι απαραίτητο κάθε νευρώνας ενός επιπέδου να είναι συνδεδεμένος σε όλους τους νευρώνες ενός άλλου επιπέδου
- **Πρόσθιας τροφοδότησης (Feedforward):** Δεν υπάρχουν συνδέσεις των νευρώνων ενός επιπέδου με νευρώνες προηγούμενου επιπέδου. Επιτρέπονται μόνο συνδέσεις με νευρώνες του επόμενου επιπέδου.
- **Με ανατροφοδότηση (Feedback):** Επιτρέπονται συνδέσεις των νευρώνων ενός επιπέδου με νευρώνες προηγούμενου ή του ίδιου επιπέδου.

2.5.3 Συνάρτηση Ενεργοποίησης (Activation Function)

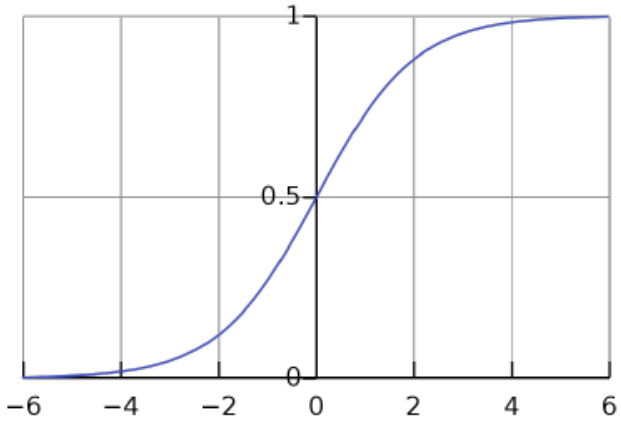
Η Συνάρτηση Ενεργοποίησης είναι βασικό κομμάτι της λειτουργίας ενός τεχνητού νευρώνα (βλ. 2.5.1). Η συνάρτηση λαμβάνει ως είσοδο το αποτέλεσμα του αθροίσματος που προκύπτει από τις τιμές εισόδων του νευρώνα και παράγει μια τιμή εξόδου η οποία προωθείται ως είσοδος στους επόμενους νευρώνες, ή τιμή εξόδου του δικτύου μας αν ο νευρώνας μας βρίσκεται στο επίπεδο εξόδου. Οι συναρτήσεις αυτές είναι εξαιρετικά χρήσιμες διότι μπορούν να οριστούν με κατάλληλο τρόπο ώστε η έξοδος τους να είναι οποιοσδήποτε αριθμός. Η λειτουργία της συνάρτησης, μπορεί να χαρακτηριστεί ως λειτουργία φίλτρου που μεταφέρει την έξοδο στο διάστημα που κρίνεται βολικότερο. Υπάρχουν διάφορες συναρτήσεις ενεργοποίησης, παρακάτω παρουσιάζονται κάποιες από τις συνηθέστερες.

1. **Σιγμοειδής Συνάρτηση**

Μετατρέπει την είσοδο της στο διάστημα (0,1). Η κανονικοποίηση των τιμών στο διάστημα αυτό οδηγεί στο να μην υπάρχουν αισθητές διαφορές των τιμών εξόδου της συνάρτησης, ένα πρόβλημα που ονομάζουμε εξασθένιση κλίσης (vanishing gradient problem).

$$f(x) = \sigma(x) = \frac{1}{1 + e^{-x}}$$

Η γραφική παράσταση της συνάρτησης φαίνεται στο *Σχήμα 26*.



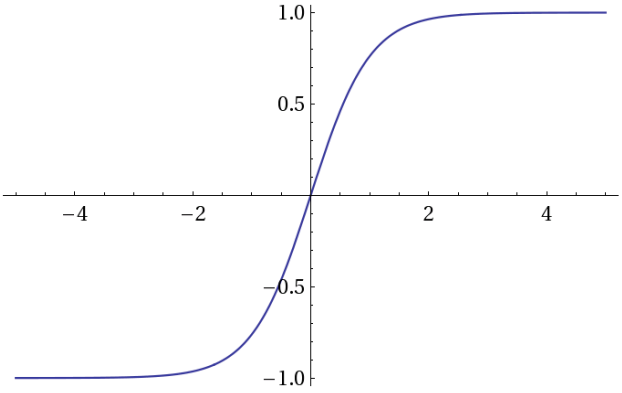
Σχήμα 26: *Σιγμοειδής Συνάρτηση, Wikipedia*

2. Συνάρτηση Υπερβολικής Εφαπτομένης

Μετατρέπει την είσοδο της στο διάστημα (-1,1). Σε αντίθεση με την σιγμοειδή, η τιμή εξόδου με χρήση της υπερβολικής συνάρτησης εφαπτομένης, παραμένει κεντραρισμένη γύρω από το 0. Όπως και η σιγμοειδής, αντιμετωπίζει το πρόβλημα εξασθένισης κλίσης.

$$f(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Η γραφική παράσταση της συνάρτησης φαίνεται στο *Σχήμα 27*.



Σχήμα 27: *Συνάρτηση Υπερβολικής Εφαπτομένης, O'Reilly*

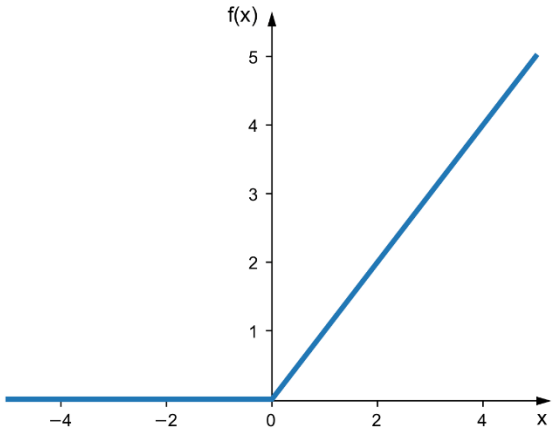
3. Μονάδα Γραμμικής Ανόρθωσης (Rectified Linear Unit – ReLU)

Η συνάρτηση αυτή είναι η πιο διαδεδομένη συνάρτηση στον χώρο των Βαθιών Νευρωνικών Δικτύων (Deep Neural Networks). Η συνάρτηση επιστρέφει τιμή 0 για τις αρνητικές εισόδους και μια γραμμική συνάρτηση για τις θετικές εισόδους. Η συνάρτηση αυτή είναι ιδανική σε

προβλήματα που δεν έχουν μεγάλες τιμές εισόδου. Επιπλέον δεν εμφανίζει το πρόβλημα της εξασθένισης κλίσης που συναντήσαμε στις δύο προηγούμενες συναρτήσεις. Ο μηδενισμός των αρνητικών τιμών οδηγεί στην αδρανοποίηση όλων των νευρώνων με αυτές τις τιμές, γεγονός που ανάλογα με τη φύση του προβλήματος, μπορεί να επηρεάσει σημαντικά τη λύση του προβλήματος.

$$f(x) = (0, \max)$$

Η γραφική παράσταση της συνάρτησης φαίνεται στο *Σχήμα 28.*



Σχήμα 28: *Μονάδα Γραμμικής Ανόρθωσης, Sebastian Raschka*

2.5.4 Συνάρτηση Κόστους (Cost Function)

Ο έλεγχος επίδοσης ενός ΤΝΔ γίνεται με τη χρήση συναρτήσεων κόστους. Για την πραγματοποίηση του ελέγχου, χρησιμοποιούμε ως δεδομένα τα δείγματα εισόδου και τη γνωστή, αναμενόμενη τιμή εξόδου του κάθε δείγματος. Η σύγκριση των αναμενόμενων τιμών με τις τιμές εξόδου που έδωσε το μοντέλο μας, οδηγεί στον υπολογισμό του σφάλματος. Με αυτό το τρόπο μπορούμε να παρακολουθήσουμε την βελτίωση της επίδοσης του δικτύου και να το τροποποιήσουμε με στόχο την ελαχιστοποίηση των σφαλμάτων που προκύπτουν. Παρακάτω παρουσιάζονται κάποιες διαδεδομένες συναρτήσεις κόστους που εφαρμόζονται στους τομείς της Στατιστικής και της Μηχανικής Μάθησης.

1. **Μέσο Τετραγωνικό Σφάλμα - Mean Squared Error (MSE):** Υπολογίζεται ο μέσος όρος των τετραγώνων των σφαλμάτων. Ο μαθηματικός τύπος είναι ο εξής:

$$J(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Pi)^2$$

όπου Y_i οι πραγματικές τιμές και P_i οι προβλέψεις.

2. **Μέσο Απόλυτο Σφάλμα - Mean Absolute Error (MAE):** Υπολογίζεται ο μέσος όρος της απόλυτης τιμής των σφαλμάτων. Ο μαθηματικός τύπος είναι ο εξής:

$$J(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}|Y_i - Pi|$$

όπου Y_i οι πραγματικές τιμές και P_i οι προβλέψεις.

3. **SVM Loss (Hinge loss):** Συναντάται σε προβλήματα κατηγοριοποίησης και χρησιμοποιείται στις Μηχανές Διανυσμάτων Υποστήριξης. Στοχεύει στο να είναι το άθροισμα των σωστών προβλέψεων μεγαλύτερο από το αντίστοιχο άθροισμα των λανθασμένων. Η συνάρτηση αυτή δεν είναι παραγωγίσιμη αλλά είναι εύκολη στον υπολογισμό της. Ο μαθηματικός τύπος είναι ο εξής:

$$J(\boldsymbol{\theta}) = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

όπου θ το διάνυσμα των παραμέτρων του εκάστοτε δικτύου.

4. **Cross Entropy Loss:** Είναι η πιο συνηθισμένη στα προβλήματα ταξινόμησης. Η λειτουργία της είναι η σύγκριση των πιθανοτικών κατανομών των προβλεπόμενων τιμών και των πραγματικών τιμών. Όσο μεγαλύτερη η απόκλιση των δύο κατανομών τόσο μεγαλύτερη είναι η τιμή της συνάρτησης. Στόχος είναι να λάβει όσο δυνατόν μικρότερες τιμές η συνάρτηση ώστε οι δύο κατανομές να ταυτιστούν. Χαρακτηριστικό αυτής της συνάρτησης είναι το γεγονός πως τοποθετεί μεγάλη ποινή σε προβλέψεις με μεγάλο βαθμό σιγουριάς (confident predictions) που όμως τελικά είναι λανθασμένες. Ο μαθηματικός τύπος είναι ο παρακάτω:

$$J(\boldsymbol{\theta}) = H(p, q) = -\sum_{x \in X} p(x)\log(q(x))$$

όπου p,q οι πιθανοτικές κατανομές των πραγματικών τιμών και των προβλεπόμενων τιμών αντίστοιχα.

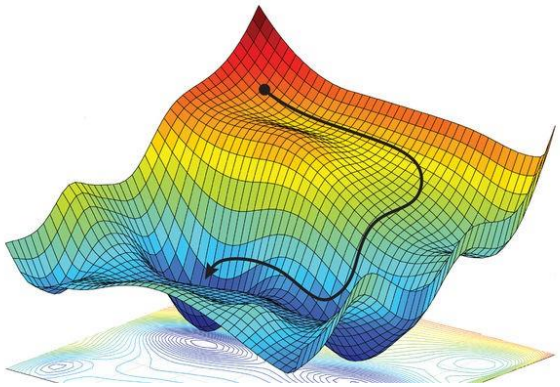
2.5.5 Αλγόριθμοι Εκπαίδευσης - Βελτιστοποίησης

Σε ένα Τεχνητό Νευρωνικό Δίκτυο ακολουθείται κάποιος αλγόριθμος εκπαίδευσης που προβλέπει την προσαρμογή των βαρών των νευρώνων με βάση συγκεκριμένο κανόνα εκπαίδευσης. Στην Επιβλεπόμενη Μάθηση, η εκπαίδευση ισοδυναμεί με ένα πρόβλημα μεταβολής των βαρών, με τρόπο ο οποίος οδηγεί στην ελαχιστοποίηση μιας συνάρτησης σφάλματος, μεταξύ πραγματικής και επιθυμητής εξόδου.

Αλγόριθμος Οπίσθιας Διάδοσης – Backpropagation

Ο κυριότερος αλγόριθμος που χρησιμοποιείται στην εκπαίδευση MLPs είναι ο αλγόριθμος οπίσθιας διάδοσης (backpropagation). Ο αλγόριθμος αυτός πρώτη φορά προτάθηκε από τον Kelley το 1960, και αργότερα το 1974 ο Werbos στη διδακτορική του διατριβή, πρότεινε την εφαρμογή του αλγορίθμου αυτού σε Τεχνητά Νευρωνικά Δίκτυα [33]. Η χρήση του σε ΤΝΔ έγινε δημοφιλής κάποια χρόνια πιο μετά όταν περιεγράφηκε εκτεταμένα το 1985 από τον Parker, και το 1986 από τους Rumelhart, Hinton and Williams [34].

Ο αλγόριθμος εφαρμόζει την ιδέα της ελαφριάς μεταβολής των βαρών σε κάθε βήμα εκπαίδευσης, ανάλογα με τη συνεισφορά τους στη συνάρτηση σφάλματος που χρησιμοποιείται, προς την κατεύθυνση που ελαχιστοποιεί την τιμή σφάλματος. Ως βήμα εκπαίδευσης θεωρείται η τροφοδότηση του μοντέλου με ένα δεδομένα ή μια μικρή ομάδα των δεδομένων (batch) και προσαρμογή των βαρών σύμφωνα με τον κανόνα εκπαίδευσης.



Σχήμα 29: *Οπτικοποίηση Κατάβασης Κλίσης, acoldbrew*

Ο γενικός κανόνας μάθησης με βάση τον backpropagation, ορίζει ότι η μεταβολή βάρους στο βήμα εκπαίδευσης είναι το γινόμενο της παραγώγου συνάρτησης σφάλματος ως προς το βάρος επί μία αριθμητική σταθερά η οποία καλείται ρυθμός μάθησης (learning rate). Η μέθοδος αυτή είναι γνωστή ως κατάβαση κλίσης (gradient descent) [30].

Αλγόριθμοι Βελτιστοποίησης – Optimization Algorithms

Οι Αλγόριθμοι Βελτιστοποίησης χρησιμοποιούνται με στόχο τη μεγιστοποίηση ή ελαχιστοποίηση μιας συνάρτησης, εν προκειμένω, την ελαχιστοποίηση της **Συνάρτησης Κόστους**. Εφαρμόζουμε τέτοιους αλγόριθμους για τον υπολογισμό και ενημέρωση των βέλτιστων τιμών των παραμέτρων του μοντέλου. Οι βασικές μετρικές που καθορίζουν την αποτελεσματικότητα ενός αλγορίθμου βελτιστοποίησης είναι η ταχύτητα σύγκλισης (πόσο γρήγορη είναι η διαδικασία εύρεσης του ελαχίστου) και η δυνατότητα γενίκευσης (δηλαδή η αποδοτικότητα του μοντέλου σε νέα δεδομένα. Η Κατάβαση Κλίσης που έχει προαναφερθεί είναι μια τέτοια τεχνική. Μερικοί τέτοιοι αλγόριθμοι, είναι ενσωματωμένοι στις βιβλιοθήκες που χρησιμοποιούνται στο πειραματικό μέρος. Γίνεται μία απλή ονομαστική αναφορά στον αλγόριθμο προσαρμοζόμενης εκτίμησης ροπής (Adaptive Moment Estimation – Adam) , τον αλγόριθμο στοχαστικής κατάβασης κλίσης (stochastic gradient descent – SGD) και τον αλγόριθμο Broyden–Fletcher–Goldfarb–Shanno (BFGS).

2.6 Μετρικές Αξιολόγησης – Evaluation Metrics

Οι Μετρικές Αξιολόγησης είναι σημαντικές για την εξαγωγή συμπερασμάτων μετά την εκπαίδευση των μοντέλων μας προκειμένου να είναι δυνατή η σύγκριση μεταξύ των μοντέλων αλλά και ως γενική ένδειξη της αποτελεσματικότητας ενός μοντέλου. Υπάρχουν διάφορες μετρικές και η επιλογή της κατάλληλης εξαρτάται πάντα από τη φύση του προβλήματος που αντιμετωπίζεται. Οι μετρικές αυτές εφαρμόζονται στα δεδομένα αξιολόγησης (test set) (βλ. 5.1). Πριν αναφερθούμε στις μετρικές που χρησιμοποιούνται πρέπει να ορίσουμε κάποιες κλάσεις δειγμάτων που προκύπτουν από τις προβλέψεις μας.

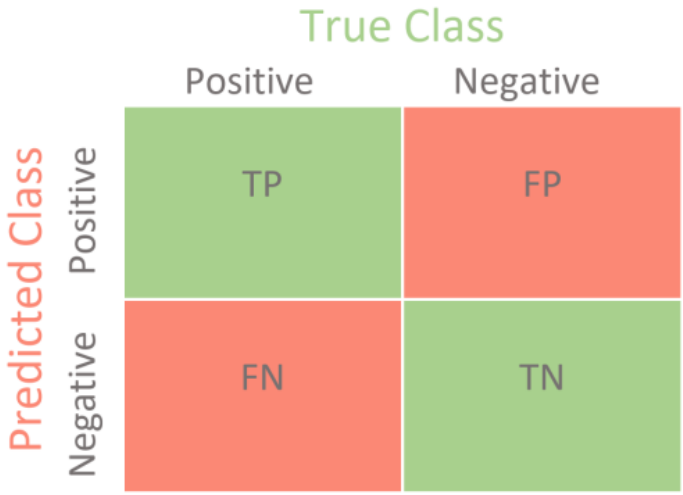
True Positive – TP: Αληθώς θετική ονομάζεται μια πρόβλεψη που εκτιμήθηκε ότι ανήκει σε μία συγκεκριμένη κλάση και αυτό ισχύει.

False Positive – FP: Ψευδώς θετική ονομάζεται μία πρόβλεψη που εκτιμήθηκε ότι ανήκει σε μία συγκεκριμένη κλάση, ενώ αυτό δεν ισχύει.

True Negative – TN: Αληθώς αρνητική ονομάζεται μια πρόβλεψη που εκτιμήθηκε ότι δεν ανήκει σε μία κλάση και αυτό ισχύει.

False Negative – FN: Ψευδώς αρνητική ονομάζεται μια πρόβλεψη που εκτιμήθηκε ότι δεν ανήκει σε μία κλάση, ενώ αυτό δεν ισχύει.

Ο συνδυασμός όλων αυτών των κλάσεων ονομάζεται Πίνακας Σύγχυσης (Confusion Matrix) και φαίνεται στο παρακάτω σχήμα:



Σχήμα 30: *Πίνακας Σύγχυσης,* [*Bryan Shalloway*]

Η οπτικοποίηση των αποτελεσμάτων με τη χρήση του πίνακα, βοηθάει στην ευκολότερη κατανόηση των αποτελεσμάτων ταξινόμησης του μοντέλου μας. Οι ορισμοί των μετρικών αξιολόγησης, είναι οι εξής:

Ορθότητα - Accuracy :

Εκφράζει πόσο ακριβής είναι η πρόβλεψη του μοντέλου σε σχέση με τα πραγματικά δεδομένα. Δηλαδή, το ποσοστό επιτυχίας του μοντέλου στην ταξινόμηση των δειγμάτων στις σωστές κατηγορίες από ολόκληρο το σύνολο των δεδομένων.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Ακρίβεια – Precision:

Είναι ο λόγος των σωστών αποτελεσμάτων πρόβλεψης μιας κλάσης προς τον συνολικό αριθμό των προβλέψεων αυτής της κλάσης. Η μετρική αυτή συνοψίζει την ικανότητα του μοντέλου να επιστρέφει ως αποτελέσματα δείγματα συναφή με την συγκεκριμένη κλάση.

$$Precision = \frac{TP}{TP + FP}$$

Ανάκληση - Recall:

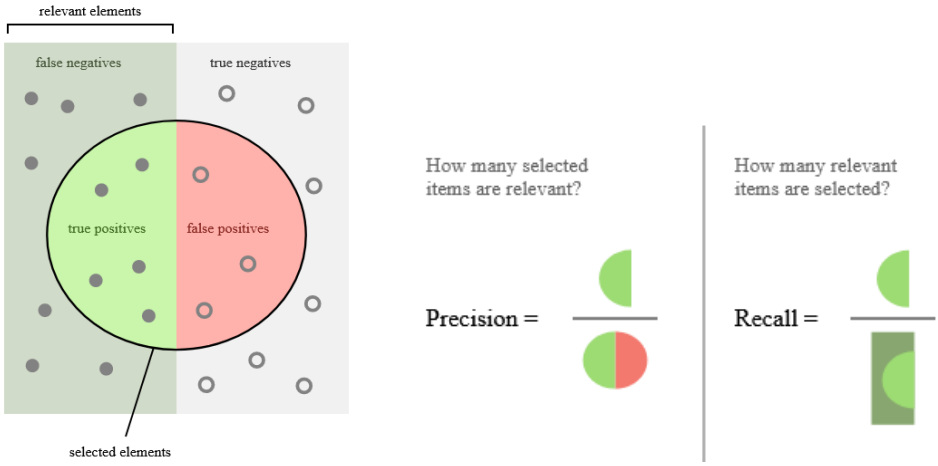
Είναι ο λόγος των σωστών προβλέψεων μιας κλάσης προς το σύνολο των δειγμάτων που πραγματικά ανήκουν σε αυτήν.

$$Recall = \frac{TP}{TP + FN}$$

F1 Score:

Αποτελεί τον αρμονικό μέσο των Precision και Recall.

$$F1 - score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$



Σχήμα 31: *Precision και Recall, Wikipedia*

Σε προβλήματα πολλών κλάσεων (multi-class classification problems) οι μετρικές Precision και Recall προκύπτουν είτε ως ο μέσος όρος των μετρικών κάθε κλάσης (macro-average), είτε ως κλάσμα με αριθμητή το άθροισμα των αριθμητών και παρονομαστή το άθροισμα των παρονομαστών των μετρικών κάθε κλάσης (micro-average).

Κεφάλαιο 3

ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ

3.1 Περιγραφή

Στα πλαίσια διεξαγωγής της παρούσας εργασίας ήταν απαραίτητη η συλλογή, καθαρισμός και προεπεξεργασία δεδομένων που αφορούν στοιχεία πτήσεων αλλά και μετεωρολογικά στοιχεία. Τα δεδομένα πτήσεων διατίθενται ανοικτά στο διαδίκτυο από το Bureau of Transportation Statistics (BTS) του Υπουργείου Μεταφορών των ΗΠΑ [3]. Ο σκοπός του BTS είναι η συλλογή, σύνθεση και ανάλυση δεδομένων από τον τομέα των μεταφορών και η διάθεση τους στο κοινό. Τα μετεωρολογικά δεδομένα που χρησιμοποιήθηκαν διατίθενται επίσης ανοιχτά για το κοινό από το National Centers for Environmental Information των ΗΠΑ [35].

Για τη σύνθεση του συνόλου δεδομένων (Dataset) χρησιμοποιήθηκαν τα δεδομένα «Reporting Carrier On-Time Performance» . Το σύνολο δεδομένων αυτό αποτελείται από δεδομένα που παρέχουν οι αερομεταφορείς και αφορούν μόνο εσωτερικές (domestic) πτήσεις εντός των ΗΠΑ. Συγκεκριμένα, όσοι αερομεταφορείς ευθύνονται για τουλάχιστον 1% των εσόδων από εσωτερικές πτήσεις, υποχρεούνται να καταβάλλουν μηνιαία δεδομένα σχετικά με τις πτήσεις τους και την επίδοση τους ως προς τις καθυστερήσεις των πτήσεων τους. Τα δεδομένα περιλαμβάνουν στοιχεία για τις αφίξεις και αναχωρήσεις εσωτερικών πτήσεων. Αναφέρουν προγραμματισμένο και πραγματικό χρόνο αναχώρησης και άφιξης, πτήσεις που ακυρώθηκαν ή εκτράπηκαν από τον προορισμό τους, χρόνο που περνάει το αεροσκάφος στο έδαφος (taxi time), αιτίες καθυστέρησης, χρόνο στον αέρα (air time) και απόσταση. Η έμφαση των δεδομένων αυτών είναι οι λεπτομέρειες γύρω από τις συνθήκες καθυστέρησης μιας πτήσης. Έτσι οι καθυστερήσεις αναφέρονται χωρισμένες σε κατηγορίες. Οι κατηγορίες καθυστέρησης διακρίνονται σε πέντε σύμφωνα με τον ορισμό τους από το Υπουργείο Μεταφορών ΗΠΑ.

1. **Carrier Delay:**

Η αιτία της ακύρωσης ή καθυστέρησης οφείλεται σε παράγοντες που βρίσκονται υπό τον έλεγχο του αερομεταφορέα (π.χ. προβλήματα συντήρησης ή πληρώματος, καθαρισμός αεροσκαφών, φόρτωση αποσκευών, τροφοδοσία κ.λπ.).

2. **Weather Delay:**

Ακραίες μετεωρολογικές συνθήκες (πραγματικές ή προβλέψεις) που κατά την κρίση του αερομεταφορέα, καθυστερούν ή αποτρέπουν τη λειτουργία πτήσης όπως ανεμοστρόβιλοι, χιονοθύελλες ή τυφώνες.

3. **National Aviation System Delay (NAS):**

Καθυστερήσεις και ακυρώσεις που οφείλονται στο εθνικό αεροπορικό σύστημα που αναφέρονται σε ένα ευρύ φάσμα συνθηκών, όπως μη ακραίες καιρικές συνθήκες, λειτουργίες αεροδρομίου, βαριά κυκλοφορία και έλεγχος εναέριας κυκλοφορίας.

4. **Security Delay:**

Καθυστερήσεις ή ακυρώσεις που προκαλούνται από εκκένωση τερματικού σταθμού ή χώρου του αεροδρομίου, επανεπιβίβαση σε αεροσκάφος λόγω παραβίασης ασφάλειας, μη λειτουργικού εξοπλισμού ελέγχου και / ή μεγάλων γραμμών που υπερβαίνουν τα 29 λεπτά σε περιοχές ελέγχου.

5. **Late Aircraft Delay:**

Μια προηγούμενη πτήση με το ίδιο αεροσκάφος αφίχθει αργά, με αποτέλεσμα την αργοπορημένη αναχώρηση της παρούσας πτήσης.

Για τα μετεωρολογικά δεδομένα χρησιμοποιήθηκε το Integrated Surface Dataset (ISD). Το ISD απαρτίζεται από παρατηρήσεις αισθητήρων στο επίπεδο της θάλασσας σε 35000 σταθμούς . Η πλειοψηφία των αισθητήρων βρίσκεται στην Βόρεια Αμερική, την Ευρώπη, την Αυστραλία και μέρη της Ασίας. Μετρήσεις που περιλαμβάνονται είναι: ποιότητα αέρα, ατμοσφαιρική πίεση, ατμοσφαιρική θερμοκρασία, βροχόπτωση, ορατότητα, θαλάσσια κύματα, παλίρροιες και άλλα. Το ISD διαθέτει επίσης ωριαία δεδομένα για ευκολία του χρήστη. Σε ορισμένους σταθμούς τα δεδομένα χρονολογούνται μέχρι και το 1901. Τη στιγμή της συγγραφής του παρόντος υπάρχουν περισσότεροι από 14000 σταθμοί των οποίων τα δεδομένα ενημερώνονται καθημερινά.

Σημειώνεται, πως δεδομένα διατίθενται σε αρχεία με comma separated values (.csv).

3.2 Συλλογή

Η παρούσα εργασία επικεντρώνεται συγκεκριμένα στα στοιχεία εσωτερικών πτήσεων (domestic flights) με προορισμό το John F. Kennedy International Airport (JFK) στην Νέα Υόρκη. Τα δεδομένα πτήσεων είναι διαθέσιμα για κάθε πολιτεία σε μηνιαία βάση. Για τα δεδομένα του JFK, ακολουθήθηκε η χειροκίνητη διαδικασία συλλογής μηνιαίων δεδομένων από τον Ιανουάριο του 2012 μέχρι τον Δεκέμβριο του 2019. Μέσω της πλατφόρμας δεδομένων του BTS υπάρχει η δυνατότητα επιλογής των χαρακτηριστικών που επιθυμεί ο χρήστης να συλλέξει. Επιλέγηκαν όσα χαρακτηριστικά είχαν να κάνουν με ώρα και ημερομηνία της πτήσης, αεροδρόμιο αναχώρησης, κωδικό πτήσης (Flight Number), μοναδικό κωδικό αεροσκάφους (Tail Number), προγραμματισμένες και πραγματικές ώρες αναχώρησης, απόσταση, χρόνο στον αέρα, συνολικά λεπτά καθυστέρησης και λεπτά καθυστέρησης ανά κατηγορία καθυστέρησης.

Με σκοπό την εκτενέστερη μελέτη και προσπάθεια δημιουργίας ενός μοντέλου προβλέψεων για τις καθυστερήσεις που προκαλούνται στο JFK επιλέγηκαν τρία τυχαία αεροδρόμια εντός των ΗΠΑ τα οποία εξετάστηκαν με μεγαλύτερη λεπτομέρεια. Τα αεροδρόμια που επιλέγηκαν είναι τα San Francisco International Airport (SFO) στην Καλιφόρνια, Dallas Fort Worth International Airport (DFW) στο Τέξας και το Chicago O'Hare International Airport (ORD) στο Ιλινόι.

Για τους τέσσερις προαναφερθέντες αερολιμένες συλλέχθηκαν δεδομένα από τους κοντινότερους μετεωρολογικούς σταθμούς που διαθέτει ο NOAA (National Oceanic and Atmospheric Administration) από τα δεδομένα ISD.

JFK: Κωδικός Σταθμού - 74486094789, Γεωγραφικές Συντεταγμένες - 40.63915,-73.76401

ORD: Κωδικός Σταθμού – 72530094846, Γεωγραφικές Συντεταγμένες - 41.96019,-87.93162

SFO: Κωδικός Σταθμού – 72494023234, Γεωγραφικές Συντεταγμένες - 37.6197,-122.3647

DFW: Κωδικός Σταθμού – 72259003927, Γεωγραφικές Συντεταγμένες - 32.8978,-97.0189

Κεφάλαιο 4

ΠΕΙΡΑΜΑΤΙΚΕΣ ΜΕΘΟΔΟΙ

4.1 Εργαλεία

Η γλώσσα προγραμματισμού που χρησιμοποιήθηκε σε ολόκληρη την εργασία είναι η Python3. Η γλώσσα Python σε συνδυασμό με συγκεκριμένες βιβλιοθήκες χρησιμοποιήθηκε τόσο για την προεπεξεργασία, καθαρισμό και συγχώνευση των δεδομένων, όσο και για τη δημιουργία και εκπαίδευση μοντέλων Μηχανικής Μάθησης. Όλα τα πειράματα εκτελέστηκαν τοπικά σε υπολογιστή με χρήση του εργαλείου Jupyter Notebooks.

Η βιβλιοθήκη Pandas διαθέτει μεθόδους για πιο εύκολη διαχείριση των δεδομένων και συγχώνευση datasets. Η βιβλιοθήκη NumPy χρησιμοποιείται για πράξεις μεταξύ πινάκων. Διαθέτει πληθώρα συναρτήσεων στο πεδίο της γραμμικής άλγεβρας, μετασχηματισμούς κ.α. Η Matplotlib είναι μια εύχρηστη βιβλιοθήκη απεικόνισης (plotting library), η οποία διαθέτει συναρτήσεις απεικόνισης γραφημάτων αλλά και εικόνων. Η Scipy περιέχει επιπλέον επιστημονικές μεθόδους. Στην παρούσα εργασία χρησιμοποιήθηκε για δημιουργία τυχαίων αριθμών. Η βιβλιοθήκη Scikit-Learn είναι μία βιβλιοθήκη ανοιχτού κώδικα (open source). Η βιβλιοθήκη χρησιμοποιήθηκε σε διάφορα στάδια της εκπαίδευσης αλγορίθμων ML. Παραδείγματα της χρήσης της είναι η χρήση της train_test_split για τον διαχωρισμό των δεδομένων σε σύνολα εκπαίδευσης (train test) και ελέγχου (test set). Περιέχει μεθόδους κανονικοποίησης δεδομένων όπως ο MinMaxScaler που χρησιμοποιήθηκε σε σημεία της εργασίας. Η σημαντικότερη συνεισφορά της βιβλιοθήκης είναι βέβαια η πληθώρα των έτοιμων ταξινομητών που διαθέτει τους οποίους χρησιμοποιήσαμε και εκπαιδεύσαμε για την δημιουργία προβλέψεων για το πρόβλημα μας. Επιπλέον χρησιμοποιήθηκε η βιβλιοθήκη XGBoost και LightGBM για την υλοποίηση και εκπαίδευση των αντίστοιχων μοντέλων.

4.2 Προεπεξεργασία Δεδομένων

4.2.1 Καθαρισμός Δεδομένων (Data Cleansing)

Το πρώτο βήμα στον καθαρισμό των δεδομένων που έχουν συλλεγεί είναι να εντοπιστούν τα δεδομένα των πτήσεων εκείνων που για κάποιο απρόοπτο λόγο έχουν ακυρωθεί ή έχουν αλλάξει πορεία. Οι πτήσεις αυτές σύμφωνα με το Bureau of Transportation Statistics, αφορούν το **1.99%** και το **0.3%** των πτήσεων αντίστοιχα, του χρονικού διαστήματος που μελετούμε. Η μελέτη των πτήσεων αυτών δεν εμπίπτει στο αντικείμενο που εξετάζεται σε αυτή την εργασία, οπότε τα δεδομένα αυτά αφαιρούνται από το σύνολο των δεδομένων. Κατόπιν της αφαίρεσης, τα δείγματα εσωτερικών πτήσεων του JFK που έχουμε στη διάθεση μας είναι 842945.

Όπως ορίζει το Υπουργείο Μεταφορών των ΗΠΑ, μια πτήση θεωρείται ότι έχει καθυστέρηση εάν η πραγματική άφιξη της είναι περισσότερα από δεκαπέντε (15) λεπτά μετά την προγραμματισμένη άφιξη της (CRS Arrival Time). Επειδή ο στόχος μας είναι η ταξινόμηση των πτήσεων σε δύο κλάσεις, τις αργοπορημένες και τις μη αργοπορημένες, αξιοποιούμε την πληροφορία από το σύνολο δεδομένων αεροδρομίου και δημιουργούμε ένα χαρακτηριστικό, το “CLASS”, το οποίο παίρνει τιμές “Delayed” και “On-Time” ανάλογα με το αν έχουν ξεπεραστεί τα δεκαπέντε λεπτά καθυστέρησης άφιξης. Ως ένα πρόβλημα δυαδικής ταξινόμησης (binary classification), αυτές οι κλάσεις αντιστοιχίζονται στις τιμές 0 και 1. Η διαδικασία αυτή πραγματοποιείται στα τέσσερα σύνολα δεδομένων αεροδρομίων στα οποία εργαζόμαστε στα πλαίσια της εργασίας (JFK, SFO, ORD, DFW) .

Στη συνέχεια το σύνολο δεδομένων μας εξετάζεται ως προς την ύπαρξη απουσιαζουσών ή λανθασμένα καταχωρημένων τιμών. Όσον αφορά τα σύνολα δεδομένων αεροδρομίων, λόγω της προσεγμένης δουλειάς του BTS, δεν υπάρχουν τέτοιες τιμές οπότε δε χρειάζεται διαχείριση αυτών των τιμών. Αναφερόμενοι στα σύνολα δεδομένων από τους μετεωρολογικούς σταθμούς, υπάρχουν απουσιάζουσες τιμές λόγω τεχνικών ζητημάτων που παρουσιάζονται ανά καιρούς από τους αισθητήρες. Ωστόσο οι τιμές αυτές δεν είναι σημαντικές σε αριθμό. Η μέθοδος που θα χρησιμοποιηθεί για την διαχείριση αυτών των τιμών είναι η μέθοδος της απόδοσης (Imputing). Σε κάθε στήλη αποδίδεται στις απουσιάζουσες τιμές, η μέση τιμή (mean value) της εκάστοτε στήλης με σκοπό την αποφυγή στατιστικών ανωμαλιών. Δεν θα ήταν συνετό στο παρόν στάδιο να αφαιρεθούν οι καταχωρήσεις αυτές εντελώς, καθώς στην

πλειοψηφία των καταχωρήσεων απουσιάζει μία μόνο τιμή, π.χ. η ορατότητα, ενώ τα υπόλοιπα χαρακτηριστικά του καιρού έχουν μετρηθεί κανονικά.

Επισημαίνεται δε, ότι στα σύνολα δεδομένων των αεροδρομίων, η ώρα αναχώρησης και άφιξης δίδονται σε τοπική ώρα (local time) για κάθε αερολιμένα. Στα σύνολο μετεωρολογικών δεδομένων, δίδονται σε Coordinated Universal Time (UTC). Για να συμβαδίζουν τα δεδομένα και για την ύπαρξη κοινού σημείου αναφοράς, όλοι οι χρόνοι μετατρέπονται σε UTC. Επιπλέον για σκοπούς διευκόλυνσης πράξεων, όλοι οι χρόνοι και ημερομηνίες μετατρέπονται σε Unix Time (αριθμός δευτερολέπτων που έχουν περάσει από τα μεσάνυχτα της 1/1/1970).

Μετά τον καθαρισμό των δεδομένων, τα datasets αεροδρομίων και καιρού είναι έτοιμα για συγχώνευση. Το σύνολο δεδομένων του διεθνή αερολιμένα JFK στην παρούσα του μορφή, ονομάζουμε "**Dataset I**" για σκοπούς μελλοντικής αναφοράς στην περιγραφή των πειραμάτων.

4.2.2 Συγχώνευση Δεδομένων

Για την πιο αποτελεσματική εκπαίδευση των μοντέλων είναι αναγκαίο να είναι διαθέσιμη πληροφορία για τον καιρό. Έχοντας καθαρίσει τα σύνολα δεδομένων που αφορούν το JFK προχωράμε στη συγχώνευση τους σε ένα σύνολο δεδομένων, στο οποίο κάθε καταχώρηση φέρει πληροφορία για όλα στα στοιχεία μιας πτήσης και επιπλέον πληροφορία για τον καιρό την ώρα άφιξης. Από τα μετεωρολογικά δεδομένα επιλέγηκαν συγκεκριμένα χαρακτηριστικά που επηρεάζουν τη δυνατότητα πτήσης ενός αεροσκάφους. Περισσότερα για την επιλογή των χαρακτηριστικών στο Κεφάλαιο 4.2.3. Η συγχώνευση γίνεται στο πεδίο "Arrival Time" από το σύνολο πτήσεων και στο πεδίο "Timestamp" από το σύνολο καιρού με τη χρήση της συνάρτησης "merge_asof" της βιβλιοθήκης Pandas. Οι μετεωρολογικές μετρήσεις προστίθενται πάνω σε κάθε καταχώρηση του συνόλου πτήσεων σύμφωνα με την ώρα που λήφθηκε η μέτρηση. Η μέτρηση του καιρού που επιλέγεται είναι αυτή με τη μικρότερη χρονική απόκλιση από την ώρα άφιξης μιας πτήσης. Η πληροφορία για την ώρα άφιξης φυσικά, αφαιρείται στη συνέχεια από το σύνολο δεδομένων, καθώς είναι πληροφορία εξόδου. Το νέο σύνολο δεδομένων, αποθηκεύεται και θα αναφερόμαστε σε αυτό ως "**Dataset II**".

Παράλληλα με τα δεδομένα για τα αεροδρόμια ενδιαφέροντος, έχουν συλλεγεί χειροκίνητα τα δεδομένα πτήσεων από όλες τις εσωτερικές πτήσεις στις ΗΠΑ για κάθε μέρα από την 1/1/2012 μέχρι την 31/12/2019. Με τη βοήθεια της Python μετρήθηκε ο αριθμός των πτήσεων για κάθε μέρα ξεχωριστά στο χρονικό διάστημα αυτό. Την πληροφορία αυτή

εισάγουμε ως το χαρακτηριστικό “Volume” στο σύνολο δεδομένων Dataset II, αφού μας είναι χρήσιμη σαν μετρική συμφόρησης του Εθνικού Αεροπορικού Συστήματος (NAS) για μια συγκεκριμένη μέρα. Μέσω της βιβλιοθήκης Sci-kit Learn, χρησιμοποιούμε τη μέθοδο MinMaxScaler για την κανονικοποίηση του Volume στο διάστημα [0,1]. Με τον τρόπο αυτό ο ταξινομητής μπορεί να αξιοποιήσει αποτελεσματικότερα την πληροφορία συμφόρησης στο δίκτυο.

Καταλυτικός παράγοντας στην ανίχνευση μιας καθυστέρησης πτήσης είναι η πληροφορία για προηγούμενη καθυστέρηση του αεροσκάφους που εκτελεί το δρομολόγιο ενδιαφέροντος. Η ιχνηλάτηση των αεροσκαφών είναι εξαιρετικά δύσκολη, αλλά και ανούσια από πλευράς εκπαίδευσης των μοντέλων να γίνει σε ολόκληρο το NAS. Συγκεκριμένα στις ΗΠΑ το 2019 υπήρχαν 5080 δημόσια αεροδρόμια. Ως εκ τούτου, έγινε η τυχαία επιλογή των τριών προαναφερθέντων αερολιμένων. Με σημείο εκκίνησης αυτούς τους αερολιμένες και προορισμό το JFK δραστηριοποιούνται 14 αερομεταφορείς. Επειδή η συντριπτική πλειοψηφία των πτήσεων (87%) εκτελείται από 4 μόνο αερομεταφορείς επιλέξαμε να ασχοληθούμε μόνο με τα δεδομένα αυτών. Οι 4 εταιρείες που εξετάζουμε είναι η JetBlue Airways, η Delta Airlines, η American Airlines και η Endeavor Air. Αξίζει να σημειωθεί ότι οι πτήσεις οι οποίες κατέφθασαν στην ώρα τους στο JFK στο διάστημα 1/1/2012 – 31/12/2019 είναι 76.37% και η ίδια κατηγορία στο δικό μας δίκτυο των τεσσάρων αεροδρομίων ήταν 76.49% προσφέροντας μια καλή ένδειξη ότι τα δεδομένα μας συμβαδίζουν με τη γενική περίπτωση.

Για την ιχνηλάτηση των αεροσκαφών χρησιμοποιήθηκε ο μοναδικός αριθμός εγγραφής αεροσκάφους γνωστός και ως “Tail Number”. Αρχικά εντοπίστηκαν τα Tail Numbers όλων των αεροσκαφών που είχαν εκτελέσει πτήσεις μέσα στο χρονικό διάστημα που εξετάζουμε και αποθηκεύτηκαν σε μια λίστα. Ακολούθως, όλες οι πτήσεις με προορισμό τα αεροδρόμια SFO, ORD, DFW φιλτραρίστηκαν ώστε να παραμείνουν μόνο οι πτήσεις που εκτελέσθηκαν από αεροσκάφη που ήταν μέσα στη λίστα. Έπειτα, έγινε ταυτοποίηση της χρονικά πλησιέστερης πτήσης αυτών των αεροσκαφών με προορισμό τα αεροδρόμια αυτά, με τις πτήσεις τους που αναχωρούσαν για JFK. Με στόχο να μη δημιουργηθούν συσχετισμοί πτήσεων που στην πραγματικότητα δεν επηρεάζουν η μία την άλλη, παραδείγματος χάρη, το αεροσκάφος με κωδικό *N702TW* προσγειώνεται στο Σαν Φρανσίσκο την 01/03 και αναχωρεί για Νέα Υόρκη τις 04/03, η χρονική διαφορά της χρονικής στιγμής άφιξης σε ένα από τα τρία αεροδρόμια SFO, ORD ή DFW και της μετέπειτα αναχώρησης του προς το JFK περιορίστηκε στις **±12 ώρες**. Η συγχώνευση έγινε και πάλι με την ίδια μέθοδο πάνω στα πεδία του πραγματικού χρόνου άφιξης στα SFO, ORD ή DFW και την προγραμματισμένη ώρα αναχώρησης προς το JFK. Τελικά, το

σύνολο δεδομένων που προέκυψε περιλάμβανε πληροφορία για όλα τα στοιχεία της εξεταζόμενης πτήσης προς το JFK, την κανονικοποιημένη συμφόρηση του δικτύου NAS, μετεωρολογικά δεδομένα για το JFK, την χρονική στιγμή άφιξης σε αυτό, και δεδομένα για την αμέσως προηγούμενη πτήση του αεροσκάφους με προορισμό ένα από τα υπόλοιπα αεροδρόμια. Το σύνολο δεδομένων που προέκυψε περιλαμβάνει 57692 δείγματα.

Στο σύνολο δεδομένων που δημιουργήθηκε εφόσον έχουμε πλέον μόνο τρία αεροδρόμια αναχώρησης προσθέτουμε με τον ίδιο τρόπο που έγινε στα δεδομένα του JFK, τις μετεωρολογικές μετρήσεις για το εκάστοτε αεροδρόμιο αναχώρησης την ώρα της προγραμματισμένης αναχώρησης των πτήσεων με προορισμό το JFK. Χωρίς ακόμα να έχει ληφθεί υπόψιν η πληροφορία για ενδεχόμενη αργοπορία του προηγούμενου δρομολογίου που εκτελούσαν τα αεροσκάφη αποθηκεύουμε το σύνολο δεδομένων που περιλαμβάνει την πληροφορία που αναγράφεται στην προηγούμενη παράγραφο και θα αναφερόμαστε σε αυτό ως "**Dataset III**".

Σκοπός της σταδιακής αυτής αύξησης της διαθέσιμης πληροφορίας που διαθέτει κάθε δείγμα στα σύνολα δεδομένων που κατασκευάσαμε είναι η σωστή μελέτη της επιρροής ενός μόνο χαρακτηριστικού κάθε φορά.

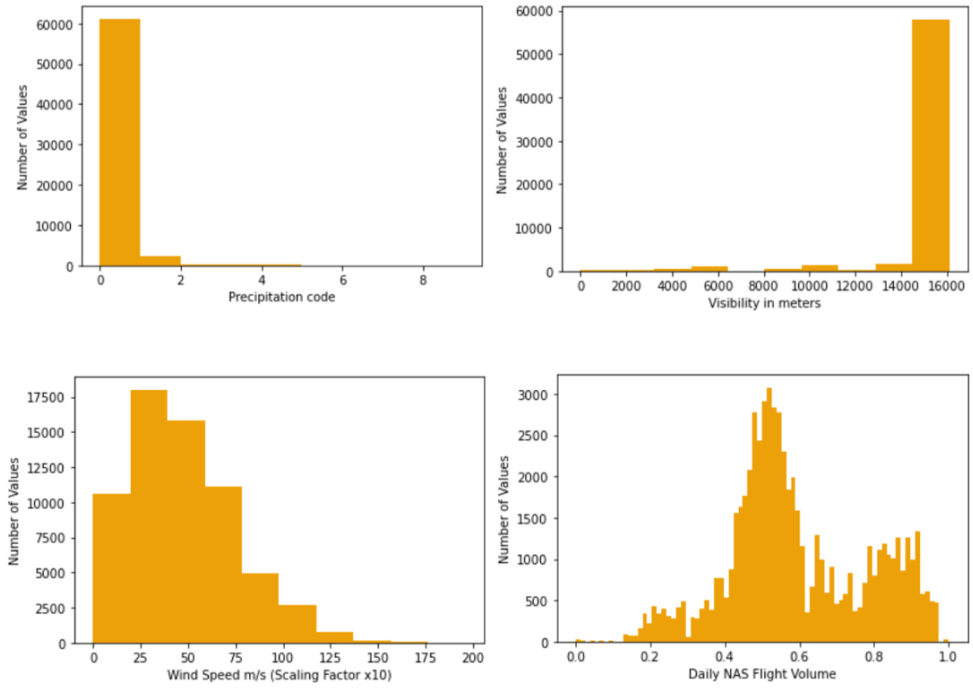
Τέλος, στο σύνολο δεδομένων μας, όπως και στο πρώτο βήμα της προεπεξεργασίας δημιουργείται παράμετρος που χαρακτηρίζει τις πτήσεις με προορισμό τα αεροδρόμια SFO, DFW, ORD σε πτήσεις με καθυστέρηση ή χωρίς. Μαζί με το χαρακτηρισμό του προηγούμενου δρομολογίου ως καθυστερημένο ή μη, ενσωματώνεται στο dataset η πληροφορία του χρόνου καθυστέρησης σε λεπτά (σε περίπτωση άφιξης πριν την προγραμματισμένη ώρα οι τιμές είναι αρνητικές), η ώρα άφιξης και το πόσο χρονικό διάστημα (slack) υπάρχει μέχρι την επόμενη αναχώρηση του αεροσκάφους σε λεπτά. Το τελικό dataset θα αναφέρεται ως "**Dataset IV**".

4.2.3 Feature Importance

Τα σύνολα δεδομένων καιρού ISD παρέχουν πολλές παραμέτρους που αφορούν τον καιρό καθώς και κωδικούς για την ποιότητα των μετρήσεων κ.α. Μετά από βιβλιογραφική μελέτη σχετικά με την επιρροή των καιρικών συνθηκών στις πτήσεις, καθώς και πειραματικές δοκιμές με τα δεδομένα οδηγούμαστε στο συμπέρασμα ότι δεν επηρεάζουν όλες οι καιρικές συνθήκες την ομαλή διεξαγωγή πτήσεων. Τα περιττά χαρακτηριστικά αφαιρέθηκαν από το σύνολο δεδομένων μας και διατηρήθηκαν μόνο όσα παρουσίασαν συσχέτιση (correlation) με την καθυστέρηση μιας πτήσης. Οι παράγοντες που επηρεάζουν έχουν αναλυθεί εκτενώς στο

Κεφάλαιο 1.4. οπότε στο παρόν γίνεται απλή αναφορά στα χαρακτηριστικά του dataset που αντιστοιχούν σε αυτούς τους παράγοντες.

- Ταχύτητα ανέμου (0-900 σε m/s, με scaling factor x10)
- Ορατότητα (0-16000 σε m, απόσταση ορατότητας στον ορίζοντα)
- Ατμοσφαιρική Θερμοκρασία (-0932 με +0618 σε Κελσίου, με scaling factor x10)
- Σημείο Υγροποίησης (-0982 με +0368 σε Κελσίου, με scaling factor x10)
- Ατμοσφαιρική Πίεση (8600-10900 σε Hectopascals, με scaling factor x10)
- Χιλιοστά Βροχής (0-9998 σε mm)
- Κωδικός Κακοκαιρίας (0-8, Κατηγορική Διατάξιμη – ordinal μεταβλητή)
- Κωδικός Έντασης Βροχής (0-9, Κατηγορική Διατάξιμη – ordinal μεταβλητή)



Σχήμα 32: *Κατανομές τιμών χαρακτηριστικών από το σύνολο δεδομένων. Πάνω αριστερά - Κωδικός έντασης βροχής, Πάνω Δεξιά – Ορατότητα, Κάτω αριστερά – Ταχύτητα Ανέμου, Κάτω Δεξιά – Πλήθος Πτήσεων κανονικοποιημένο [0,1].*

Όλα τα χαρακτηριστικά του συνόλου δεδομένων έχουν εξεταστεί ως προς τη στατιστική τους σημασία για να διαπιστωθεί εάν η πληροφορία που παρέχουν δεν είναι αναγκαία. Παραδείγματος χάρη, στη γενική περίπτωση οι πτήσεις που κατηγοριοποιούνται ως αργοπορημένες αντιπροσωπεύουν το **23,51%**, αν πάρουμε μόνο το υποσύνολο των πτήσεων στα οποία η Ορατότητα στο αεροδρόμιο αναχώρησης παίρνει τιμή κάτω από 5000m, παρατηρούμε ότι αυτές οι πτήσεις αντιπροσωπεύουν το **38,4%.** Οι πτήσεις στις οποίες η

ταχύτητα του ανέμου ξεπερνά τα 140 m/s έχουν **44,5%** καθυστερήσεις. Παρατηρούμε επίσης ότι οι πτήσεις που αναχωρούν από το αεροδρόμιο του SFO καταλήγουν καθυστερημένες με ποσοστό **20,6%,** ενώ αντίστοιχα από το DFW, έχουν ποσοστό **30,0%.** Ακόμα παρατηρούμε ότι εάν η θερμοκρασία κατεβεί στους -15 βαθμούς Κελσίου οι αργοπορημένες πτήσεις αντιστοιχούν στο **43,1%.** Επιπλέον παρατηρήθηκε ότι σε περιπτώσεις που ο όγκος πτήσεων είναι υψηλός δεν επηρεάζονται οι καθυστερήσεις σε αρκετά σημαντικό στατιστικά βαθμό, όμως εάν ο αριθμός τους είναι χαμηλός τότε είναι λιγότερο πιθανές. Συγκεκριμένα, για μέρες με όγκο κάτω των 13000 πτήσεων, το ποσοστό καθυστέρησης ανέρχεται στα **15,8%** σε αντίθεση με το 23,51% της γενικής περίπτωσης.

4.2.4 Τεχνικές Προεπεξεργασίας

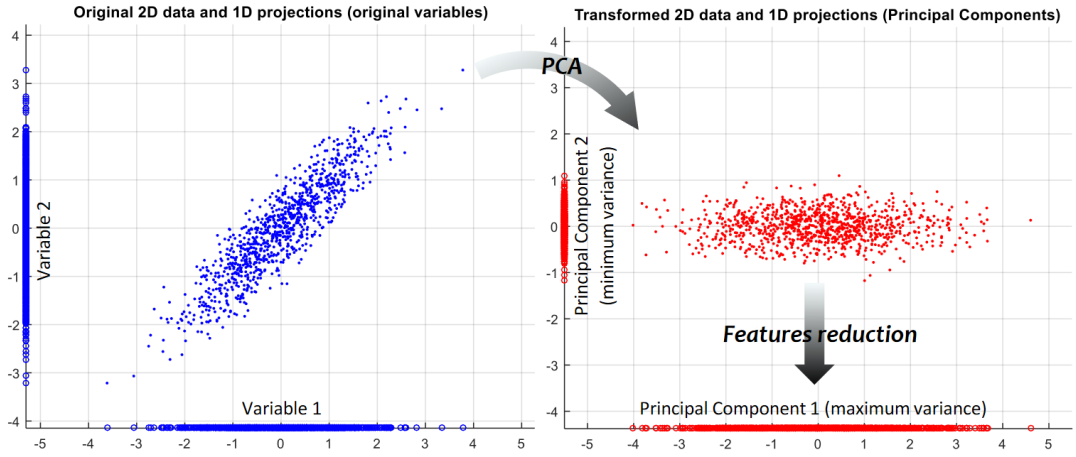
4.2.4.1 Κατώφλι Διακύμανσης (Variance Threshold)

Είναι μια απλή τεχνική για αφαίρεση μη σημαντικών χαρακτηριστικών. Η υλοποίηση και εφαρμογή της τεχνικής έγινε μέσω της βιβλιοθήκης Sci-Kit Learn. Βασίζεται στην ιδέα ότι όταν οι τιμές που λαμβάνει ένα χαρακτηριστικό έχουν σχετικά μικρή διακύμανση σε όλα τα στοιχεία ενός dataset τότε δε μπορεί να αξιοποιηθεί για την εύρεση χρήσιμων μοτίβων και άρα μπορεί να αφαιρεθεί από τα δεδομένα. Δεδομένου ενός ορίου (κατώφλι) διακύμανσης αφαιρούνται, λοιπόν, τα χαρακτηριστικά που έχουν διακύμανση μικρότερη αυτού. Η συγκεκριμένη τεχνική εμπίπτει στη κατηγορία των μεθόδων διήθησης. Για τους σκοπούς των πειραμάτων μας μελετήθηκε η διακύμανση σε διάφορες τιμές ωστόσο δε βρέθηκε κάποια τιμή που να συνεισφέρει στην ικανότητα εκπαίδευσης των μοντέλων μας. Ορίστηκε στην τιμή μηδέν για να απομακρύνει τα χαρακτηριστικά που είναι σταθερά, όπως π.χ. ο κωδικός του αεροδρομίου άφιξης (JFK).

4.2.4.2 Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis)

Μέχρι στιγμής έχουμε ασχοληθεί με την αφαίρεση χαρακτηριστικών που δεν προσφέρουν σημαντική πληροφορία κάνοντας επιλογή μεταβλητών (feature selection). Εν συνεχεία θα κάνουμε εξαγωγή νέων χαρακτηριστικών (feature extraction) σε ένα χώρο μικρότερων διαστάσεων. Μία βασική τεχνική feature extraction είναι η ανάλυση σε κύριες συνιστώσες (PCA). Η PCA είναι μετασχηματισμός δεδομένων, που ενδεχομένως αποτελούνται

από συσχετισμένες μεταβλητές, με τον οποίο τα νέα δεδομένα αποτελούνται από νέες ασυσχέτιστες μεταβλητές με μέγιστη διακύμανση. Με την τεχνική αυτή επιδιώκουμε μεγιστοποίηση της διακύμανσης, ελαχιστοποίηση της συσχέτισης και ελάττωση διαστάσεων [36] .



Σχήμα 33: *Μετασχηματισμός PCA, I. Kalatzis*

Ο μετασχηματισμός PCA γίνεται με προβολή των σημείων σε μια διεύθυνση στην οποία έχουμε μεγιστοποίηση της διακύμανσης (μέγιστο εύρος). Ταυτόχρονα ελαχιστοποιείται το άθροισμα των αποστάσεων των σημείων από την ευθεία. Κατά το μετασχηματισμό οδηγούμε ταυτόχρονα τις νέες μεταβλητές να είναι ασυσχέτιστες.

Ως ασυσχέτιστες, ορίζονται δύο μεταβλητές όταν η συνδιακύμανση (covariance) τους είναι μηδέν. Η πρακτική σημασία αυτού είναι ότι η μεταβολή της μίας μεταβλητής δεν οδηγεί σε μεταβολή της άλλης.

Ο μετασχηματισμός PCA επιτυγχάνεται με περιστροφή των αρχικών δεδομένων μέχρι το νέο σύστημα συντεταγμένων να γίνει αυτό των πρωτευόντων αξόνων. Έτσι, μία μεταβλητή αποκτά μεγάλη διακύμανση (πρωτεύοντας άξονας 1), ενώ γύρω από αυτήν (στις υπόλοιπες μεταβλητές, δηλ. στον πρωτεύοντα άξονα 2 και τυχόν υπόλοιπους) έχουμε μικρή διακύμανση. Η ελάττωση του πλήθους των διαστάσεων επιτυγχάνεται μετά την προβολή των δεδομένων σε χώρο λιγότερων διαστάσεων μετά το μετασχηματισμό. Το μεγαλύτερο μέρος της χρήσιμης πληροφορίας των μεταβλητών διατηρείται [37].

4.2.4.3 Υπερδειγματοληψία και Υποδειγματοληψία

Σε ένα σύνολο δεδομένων όταν τα πλήθη δειγμάτων των κλάσεων διαφέρουν σημαντικά μεταξύ τους χρησιμοποιούμε τον όρο Μη-Ισορροπημένο (imbalanced) Dataset. Αν και δεν υπάρχει κάποιος ορισμός, εμπειρικά όταν ο λόγος μεταξύ του αριθμού των δειγμάτων σε ένα σύνολο δεδομένων δύο κλάσεων ξεπερνά τα 2:3 θεωρούμε το σύνολο μη ισορροπημένο. Οι περισσότεροι ταξινομητές τείνουν να εκπαιδεύονται πιο αποτελεσματικά εάν το πλήθος των δειγμάτων στις διαφορετικές κλάσεις είναι ισορροπημένο.

Στο σύνολο δεδομένων που έχουμε κατασκευάσει ο λόγος της κλάσης “On-Time” με της κλάσης “Delayed” είναι 3,25:1. Αυτό σημαίνει ότι το σύνολο μας δεν είναι ισορροπημένο. Παρόλα αυτά το σύνολο μας αποτελείται από μεγάλο αριθμό δειγμάτων (57692 δείγματα) με αποτέλεσμα η ισορροπία των κλάσεων στη συγκεκριμένη περίπτωση να μην είναι καθοριστικής σημασίας αφού υπάρχουν αρκετά δεδομένα για αποτελεσματική εκπαίδευση.

Υπάρχουν δύο μέθοδοι για να εξισορροπήσουμε τις κλάσεις σε ένα σύνολο δεδομένων. Στη μέθοδο της υπέρδειγματοληψίας (oversampling) επιλέγονται τυχαία δείγματα από την κλάση που υστερεί σε πλήθος δειγμάτων και επαναλαμβάνονται μέχρι να εξισωθεί ο αριθμός των δειγμάτων μεταξύ των κλάσεων. Αντίθετα, στην υποδειγματοληψία (undesrampling), επιλέγεται η κλάση με τα περισσότερα δείγματα και αφαιρούνται από αυτήν δείγματα μέχρι να φτάσουν το πλήθος των δειγμάτων της άλλης κλάσης.

Στα πλαίσια της πειραματικής διαδικασίας χρησιμοποιήθηκαν οι μέθοδοι της βιβλιοθήκες imbalanced learn, RandomOverSampler και RandomUnderSampler αντίστοιχα. Παρατηρήθηκε ότι με τη χρήση και των δύο μεθόδων τα μοντέλα μας εκπαιδεύονταν λιγότερο αποτελεσματικά και είχαμε μια πτώση της ορθότητας (accuracy) της τάξης του 1%.

4.2.4.4 Κωδικοποίηση One-Hot

Για κατηγορικές μεταβλητές η κωδικοποίηση σε ακέραιους δεν είναι αρκετή. Σε αυτή την περίπτωση χρησιμοποιούμε κωδικοποίηση One-Hot κατά την οποία αφαιρείται η κατηγορική μεταβλητή και προστίθεται μια νέα δυαδική μεταβλητή που αναφέρεται σε κάθε μοναδική τιμή της μεταβλητής αυτής. Στην εργασία αυτή τέτοια κωδικοποίηση χρησιμοποιήθηκε για τους αερομεταφορείς και για τους αερολιμένες.

Human-Readable Machine-Readable

Pet
Cat
Dog
Turtle
Fish
Cat

Cat	Dog	Turtle	Fish
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
1	0	0	0

Σχήμα 34: *Παράδειγμα κωδικοποίησης One-Hot, Morioh*

Κεφάλαιο 5

ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Σε κάθε έκφανση του συνόλου δεδομένων (Dataset I-IV) μας για πρόβλεψη καθυστέρησης πτήσεων έγινε εκπαίδευση μοντέλων η οποία αξιολογήθηκε σύμφωνα με τις μετρικές που αναφέρονται στο Κεφάλαιο 2.6. Σε κάθε μοντέλο για την παραγωγή των βέλτιστων δυνατών αποτελεσμάτων έπρεπε να γίνει βελτιστοποίηση υπερ-παραμέτρων (hyperparameter tuning). Στο παρόν κεφάλαιο εξηγείται η διαδικασία βελτιστοποίησης των υπερ-παραμέτρων των μοντέλων, καθώς και παρατίθενται και σχολιάζονται πίνακες με τις τιμές κάθε μετρικής για κάθε μοντέλο όπως προέκυψαν στην αξιολόγηση του συνόλου ελέγχου (test set).

5.1 Βελτιστοποίηση Υπερ-Παραμέτρων με χρήση GridSearchCV

Η απόδοση όλων των πιθανών συνδυασμών υπερ-παραμέτρων στο εκάστοτε μοντέλο γίνεται με τη μέθοδο της αναζήτησης πλέγματος (Grid Search). Η αναζήτηση πλέγματος είναι απλά η εξαντλητική αναζήτηση όλων των συνδυασμών ενός ορισμένου συνόλου τιμών για κάθε υπερ-παράμετρο του μοντέλου. Οι τιμές αυτές ορίζονται χειροκίνητα βάσει δοκιμών και εμπειρικής γνώσης. Η μέθοδος πρέπει να καθοδηγείται από μια μετρική επίδοσης η οποία αποτιμάται πάνω στο σύνολο ελέγχου (test set) με τη χρήση διασταυρωμένης επικύρωσης. (Cross Validation). Η υλοποίηση της μεθόδου έγινε μέσω της συνάρτησης GridSearchCV της βιβλιοθήκης Sci-kit Learn.

Κατά την τεχνική αυτή τα δεδομένα εκπαίδευσης (training set) χωρίζονται σε ένα σταθερό αριθμό πτυχών στον οποίο θα γίνει το Cross Validation. Σε κάθε επανάληψη της τεχνικής, χρησιμοποιείται μία πτυχή των δεδομένων για εκτίμηση των αποτελεσμάτων της εκπαίδευσης και οι εναπομείναντες πτυχές σαν δεδομένα εκπαίδευσης. Χρησιμοποιώντας αυτή τη μέθοδο έχουμε αποτελεσματικότερη εκπαίδευση και αποφεύγουμε το πρόβλημα της υπερπροσαρμογής (overfitting) Κεφ. 5.2 . Σε όλα τα πειράματα χρησιμοποιήθηκε 5-Fold Cross Validation για τη διασφάλιση της σωστής εκπαίδευσης.

Για την εκπαίδευση των μοντέλων το σύνολο δεδομένων χωρίστηκε σε σύνολο εκπαίδευσης και σύνολο ελέγχου. Έγινε χρήση της μεθόδου train_test_split από τη βιβλιοθήκη Sci-kit learn. Το **σύνολο ελέγχου** αποτελεί το 30% των δεδομένων, και το υπόλοιπο 70% χρησιμοποιήθηκαν για την εκπαίδευση.

Για κάθε μοντέλο εξετάστηκαν οι ακόλουθες υπερπαραμέτροι:

- **KNN**

n_neighbours

- **Random Forest**

bootstrap, max_depth, max_features, min_samples leaf, min_samples_split, n_estimators

- **Logistic Regression**

C, max_iter, penalty, solver

- **AdaBoost**

algorithm, learning_rate, n_estimators

- **Decision Trees**

criterion, max_depth, max_features, min_samples_leaf, min_samples_split, splitter

- **XGBoost**

colsample_bytree, eta, gamma, learning_rate, max_depth, min_child_weight, subsample

- **Multi-Layer Perceptron**

activation, alpha, hidden_layer_sizes,learning_rate,solver

- **Light GBM**

colsample_bytree, min_child_samppes, min_child_weight, subsample, num_leaves, reg_alpha, reg_lambda

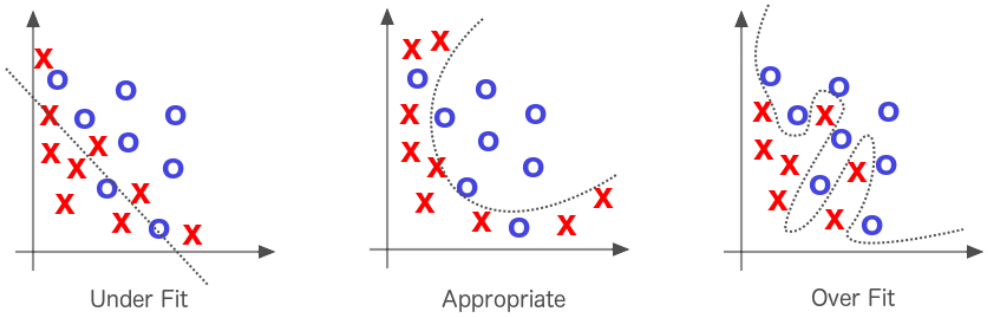
- **Support Vector Machines**

C, kernel, gamma

5.2 Υπερπροσαρμογή (Overfitting)

Ένα από τα πιο σημαντικά χαρακτηριστικά που πρέπει να διαθέτει ένα μοντέλο Μηχανικής Μάθησης, είναι να μπορεί να γενικεύσει την ικανότητα πρόβλεψης που έχει αποκτήσει από την διαδικασία εκπαίδευσης στα δεδομένα ελέγχου. Ένα συνηθισμένο

πρόβλημα που προκύπτει κατά τη διάρκεια της εκπαίδευσης είναι η υπερβολική προσαρμογή του μοντέλου στα δεδομένα εκπαίδευσης αλλά η αδυναμία του να γενικεύσει τις προβλέψεις του. Αυτό το φαινόμενο είναι γνωστό ως Υπερπροσαρμογή (Overfitting) [38]. Ένα υπερπροσαρμοσμένο μοντέλο επιτυγχάνει εξαιρετικά υψηλές επιδόσεις στο σύνολο δεδομένων εκπαίδευσης αλλά καθόλου ικανοποιητικές επιδόσεις στο σύνολο ελέγχου. Για την αποφυγή του προβλήματος αυτού εφαρμόζουμε Cross Validation. Στα αποτελέσματα που ακολουθούν σημειώνεται και η προσαρμογή των μοντέλων στα δεδομένα για να μπορούμε να εξετάσουμε εάν υπάρχει πρόβλημα υπερπροσαρμογής.



Σχήμα 35: *Απεικόνιση Προσαρμοστικότητας Μοντέλου, .py*

5.3 Αποτελέσματα

Για την αξιολόγηση των πειραμάτων χρησιμοποιήθηκαν οι μετρικές που αναφέρθηκαν στο Κεφάλαιο 2.6. Η βάση (baseline) για την αξιολόγηση των προβλέψεων στα δεδομένα μας είναι το **76,49%**. Ο λόγος είναι ότι το ποσοστό αυτό αντιστοιχεί στη μεγάλη κλάση, τις πτήσεις "On-Time", επομένως δεν υπάρχει νόημα να ασχοληθούμε με μοντέλα που προσφέρουν αποτελέσματα χαμηλότερης ορθότητας από το συγκεκριμένο ποσοστό. Σημειώνεται ότι οι μετρικές precision και recall αναφέρονται με σταθμισμένο μέσο όρο λόγω της ανισορροπίας των κλάσεων, για μια πιο ορθή εικόνα των αποτελεσμάτων.

Σημαντικά σχόλια για τα πειράματα:

Για την εξαγωγή των τελικών αποτελεσμάτων δοκιμάστηκαν πολλοί συνδυασμοί από μεθόδους προεπεξεργασίας δεδομένων και σύνολα υπερ-παραμέτρων. Πρέπει να σημειωθεί ότι μαζί με όλους τους ταξινομητές δοκιμάστηκε η προεπεξεργασία με τη χρήση της μεθόδου

Ανάλυσης σε Κύριες Συνιστώσες (PCA) αλλά σε συνδυασμό με ορισμένους ταξινομητές δεν είχαμε καλύτερα αποτελέσματα από τα αποτελέσματα χωρίς τη χρήση της μεθόδου. Τελικά, οι ταξινομητές οι οποίοι επωφελούνταν από την PCA στα πλαίσια της εργασίας, ήταν ο Decision Tree Classifier, ο SVM και το Multi-Layer Perceptron.

Επιπλέον στα SVM κρίθηκε αναγκαίο για τη σύγκλιση του ταξινομητή να οριστεί πεπερασμένος αριθμός επαναλήψεων. Η παράμετρος max_iter ορίστηκε στην τιμή 15000. Το ίδιο ισχύει και στον ταξινομητή Λογιστικής Παλινδρόμησης που ορίστηκε στην τιμη 200.

5.3.1 Αποτελέσματα I

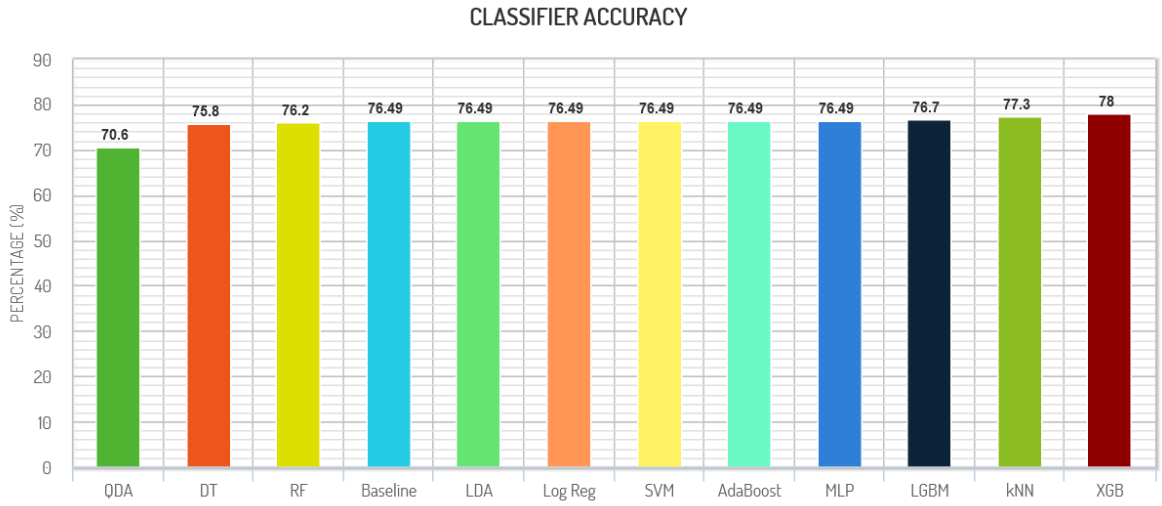
Εξετάζουμε το σύνολο δεδομένων Dataset I που περιέχει ως πληροφορία μόνο τα στοιχεία των πτήσεων με προορισμό το αεροδρόμιο JFK. Μετά το πέρας της εκπαίδευσης και της βελτιστοποίησης υπερ-παραμέτρων, τα μοντέλα μας στο σύνολο Dataset I έδωσαν τα ακόλουθα αποτελέσματα:

Μοντέλο	Precision	Recall	F1	Test Accuracy	Train Accuracy
kNN	0.74	0.77	0.75	0.773	0.777
Random Forest	0.70	0.76	0.68	0.762	0.794
Linear Discriminant Analysis	0.58	0.76	0.66	0.765	0.765
Logistic Regression	0.58	0.76	0.66	0.765	0.765
AdaBoost Classifier	0.58	0.76	0.66	0.765	0.765
Quadratic Discriminant Analysis	0.65	0.71	0.67	0.706	0.703
Decision Tree	0.66	0.76	0.67	0.758	0.770
Support Vector Machines	0.58	0.76	0.66	0.765	0.765
Light GBM	0.72	0.77	0.70	0.767	0.779

XGBoost	**0.76**	**0.78**	**0.71**	**0.780**	**0.851**
MLP	0.58	0.76	0.66	0.765	0.765

Πίνακας 2: *Αποτελέσματα Εκπαίδευσης στο Dataset I*

Όπως γίνεται αμέσως ξεκάθαρο, η πληροφορία αυτή δεν είναι αρκετή για αποτελεσματική εκπαίδευση των μοντέλων μας. Συγκεκριμένα πολλοί ταξινομητές αδυνατούν να εκπαιδευτούν γι' αυτό καταλήγουν να ταξινομούν όλα τα δείγματα στη μεγάλη κλάση και πετυχαίνουν το baseline. Η συμπεριφορά αυτή φανερώνεται στους ταξινομητές LDA, Logistic Regression, AdaBoost, SVM και MLP (76.49%). Ακόμα χειρότερα αποδίδουν οι QDA, Decision Tree Classifier και Random Forest οι οποίοι παρόλο που δεν ταξινόμησαν με αυτό τον τρόπο τα δεδομένα η εκπαίδευση τους οδήγησε σε χειρότερη επίδοση από αυτή της βάσης, με τον χείριστο να είναι ο Quadratic Discriminant Analysis. Ο Light GBM οριακά έχει καλύτερη επίδοση από τη βάση, αποτέλεσμα αμελητέο. Ελάχιστα καλύτερη επίδοση στο σύνολο ελέγχου παρουσιάζει ο KNN αφού αν και πολύ απλός ταξινομητής αποδίδει 0.8% πάνω από το baseline. Όσον αφορά την ορθότητα, καλύτερος από όλους φαίνεται να είναι ο αλγόριθμος ενίσχυσης κλίσης XGBoost (78%) ο οποίος αποδίδει 1.5% πάνω από το baseline στο σύνολο ελέγχου παρόλο που παρατηρούμε το φαινόμενο της υπερπροσαρμογής στο σύνολο εκπαίδευσης.



Σχήμα 36: *Γραφική αναπαράσταση ορθότητας ταξινομητών για το σύνολο ελέγχου του Dataset I*

5.3.2 Αποτελέσματα ΙΙ

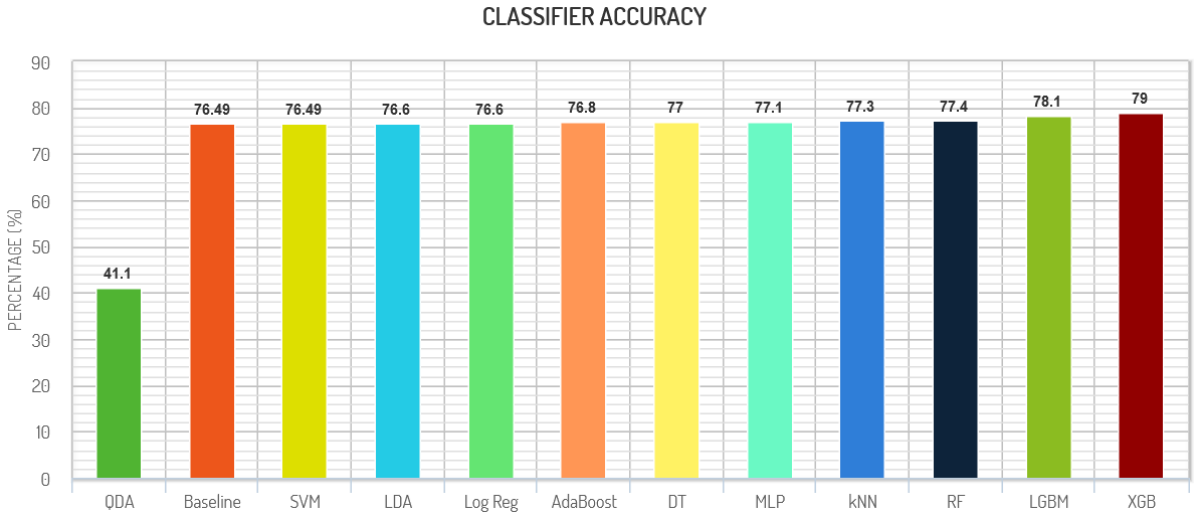
Εξετάζουμε το σύνολο δεδομένων Dataset II που περιέχει ως πληροφορία τα στοιχεία των πτήσεων με προορισμό το αεροδρόμιο JFK και τα μετεωρολογικά δεδομένα σε αυτό κατά την ώρα της προσγείωσης. Μετά το πέρας της εκπαίδευσης και της βελτιστοποίησης υπερ-παραμέτρων, τα μοντέλα μας στο σύνολο Dataset II έδωσαν τα ακόλουθα αποτελέσματα:

Μοντέλο	**Precision**	**Recall**	**F1**	**Test Accuracy**	**Train Accuracy**
kNN	0.74	0.77	0.75	0.773	0.777
Random Forest	0.74	0.77	0.71	0.774	0.823
Linear Discriminant Analysis	0.72	0.77	0.69	0.766	0.769
Logistic Regression	0.72	0.77	0.68	0.766	0.769
AdaBoost Classifier	0.73	0.77	0.68	0.768	0.770
Quadratic Discriminant Analysis	0.67	0.41	0.43	0.411	0.408
Decision Tree	0.74	0.77	0.71	0.770	0.790
Support Vector Machines	0.58	0.76	0.66	0.765	0.765
Light GBM	0.75	0.78	**0.75**	0.781	0.839
XGBoost	**0.77**	**0.79**	0.74	**0.790**	**0.933**
MLP	0.74	0.77	0.72	0.771	0.800

***Πίνακας 3**: Αποτελέσματα Εκπαίδευσης στο Dataset II*

Η προσθήκη πληροφορίας για τις καιρικές συνθήκες στο JFK φαίνεται πως είναι χρήσιμη εφόσον πλέον οι περισσότεροι ταξινομητές αποκτούν τη δυνατότητα εκπαίδευσης και σχεδόν όλοι παρουσιάζουν βελτίωση. Την εξαίρεση αποτελεί ο QDA ο οποίος παρόλο που ήδη βρισκόταν κάτω από το baseline παρουσιάζει τρομερή πτώση στην επίδοση του και επιτυγχάνει accuracy μόνο 41.1%. Αποτυχία εκπαίδευσης εξακολουθεί να υπάρχει στον SVM ο οποίος

συνεχίζει να ταξινομεί όλα τα δείγματα στη μεγάλη κλάση “On-Time”. Οι υπόλοιποι ταξινομητές έχουν όλοι καλύτερη επίδοση σε αυτό το Dataset απ’ ότι προηγουμένως. Συμπεραίνουμε ότι η επιπλέον πληροφορία σίγουρα είναι επωφελής στην εκπαίδευση. Καλύτερα από το Baseline είναι πλέον ο Decision Tree Classifier (77%) και ο Random Forest (77.4%). Ο ταξινομητής K-Nearest Neighbors είναι σταθερός, αποτέλεσμα αναμενόμενο αφού οι πτήσεις που μελετάμε είναι ακριβώς οι ίδιες. Παρατηρούμε ότι οι ταξινομητές LDA, Logistic Regression, AdaBoost και MLP δεν ταξινομούν πλέον όλα τα δείγματα του συνόλου ελέγχου στην κλάση “On-Time” και έχουν όλοι από ελάχιστη έως μικρή βελτίωση πάνω από το baseline. Η μεγαλύτερη αύξηση στο accuracy από τα προηγούμενα αποτελέσματα (+1.4%) παρουσιάζεται στον Light GBM ο οποίος έχει επίδοση 78.1%. Την καλύτερη επίδοση συνεχίζει να έχει ο XGBoost ο οποίος φτάνει το **79%**, 2.5% πάνω από το baseline.



Σχήμα 37: *Γραφική αναπαράσταση ορθότητας ταξινομητών για το σύνολο ελέγχου του Dataset II*

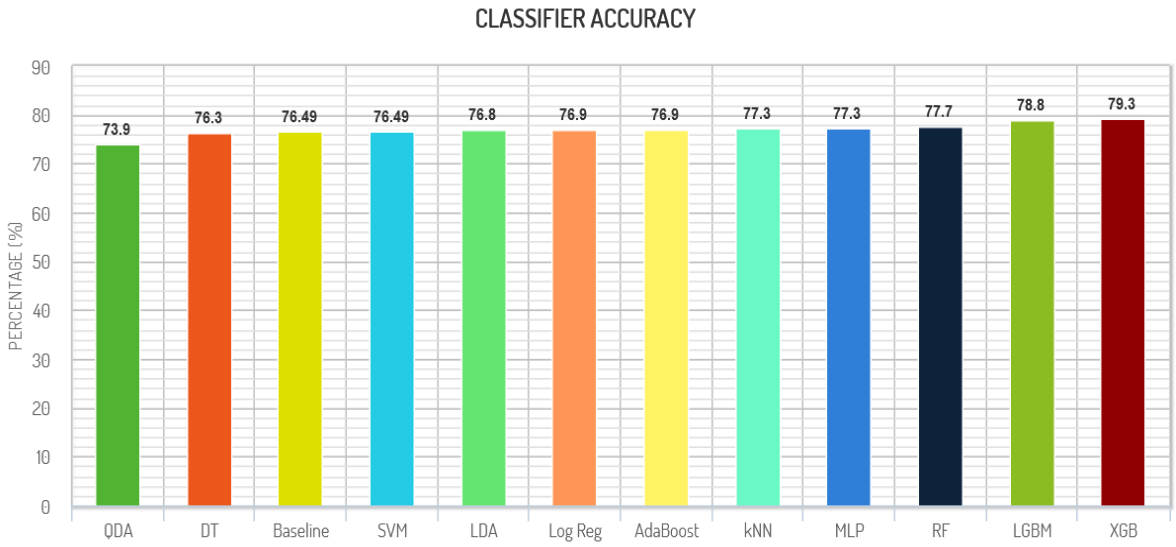
5.3.3 Αποτελέσματα III

Εξετάζουμε το σύνολο δεδομένων Dataset III που περιέχει ως πληροφορία τα στοιχεία των πτήσεων με προορισμό το αεροδρόμιο JFK και τα μετεωρολογικά δεδομένα σε αυτό κατά την ώρα της προσγείωσης. Περιλαμβάνει επιπλέον τον όγκο όλων των πτήσεων ημερήσια στο NAS και τον καιρό στο αεροδρόμιο αναχώρησης. Μετά το πέρας της εκπαίδευσης και της βελτιστοποίησης υπερ-παραμέτρων, τα μοντέλα μας στο σύνολο Dataset III έδωσαν τα ακόλουθα αποτελέσματα:

Μοντέλο	Precision	Recall	F1	Test Accuracy	Train Accuracy
kNN	0.74	0.77	0.75	0.773	0.777
Random Forest	0.76	0.78	0.71	0.777	0.827
Linear Discriminant Analysis	0.72	0.77	0.70	0.768	0.771
Logistic Regression	0.72	0.77	0.70	0.769	0.771
AdaBoost Classifier	0.74	0.77	0.69	0.769	0.773
Quadratic Discriminant Analysis	0.70	0.74	0.71	0.739	0.745
Decision Tree	0.71	0.76	0.69	0.763	0.782
Support Vector Machines	0.58	0.76	0.66	0.765	0.765
Light GBM	0.76	0.79	**0.76**	0.788	0.886
XGBoost	**0.77**	**0.79**	0.75	**0.793**	**0.961**
MLP	0.74	0.77	0.73	0.773	0.801

***Πίνακας 4**: Αποτελέσματα Εκπαίδευσης στο Dataset III*

Η προσθήκη νέων μεταβλητών στο σύνολο ανεβάζει τις επιδόσεις των ταξινομητών μας. Η αύξηση δεν είναι τόσο μεγάλη όσο η διαφορά που είχε παρουσιαστεί στα Αποτελέσματα ΙΙ από τα Αποτελέσματα Ι, όμως είναι εμφανές ότι επίδρασε θετικά. Σταθερός παραμένει ο KNN αφού οι πτήσεις που εξετάζονται είναι οι ίδιες. Το ίδιο παρατηρείται και στον SVM, η εκπαίδευση του οποίου δεν έχει αποτέλεσμα και έτσι έχουμε ταξινόμηση όλων των δειγμάτων σε μία κλάση. Τη χειρότερη επίδοση σημειώνει για ακόμη μια φορά ο Quadratic Discriminant Analysis, όμως φαίνεται ότι οι περισσότερες μεταβλητές βοήθησαν στην επίδοση του αφού παρουσιάζει τεράστια αύξηση, από 41.1% σε 73.9%, εξακολουθεί όμως να βρίσκεται κάτω από τη βάση. Ο Decision Tree Classifier δεν ταξινομεί σωστά πλέον τα δεδομένα και η επίδοση του είναι για ακόμη μία φορά κάτω από το Baseline. Την κορυφαία τριάδα ως προς την μετρική του Accuracy συμπληρώνουν ξανά ο Random Forest Classifier, ο Light GBM και ο XGBoost. Ο κορυφαίος είναι ο XGBoost του οποίου η ορθότητα αυξάνεται κατά 0.3% και ταξινομεί το 79.3% των δειγμάτων του συνόλου ελέγχου στη σωστή κλάση.



Σχήμα 38: *Γραφική αναπαράσταση ορθότητας ταξινομητών για το σύνολο ελέγχου του Dataset III*

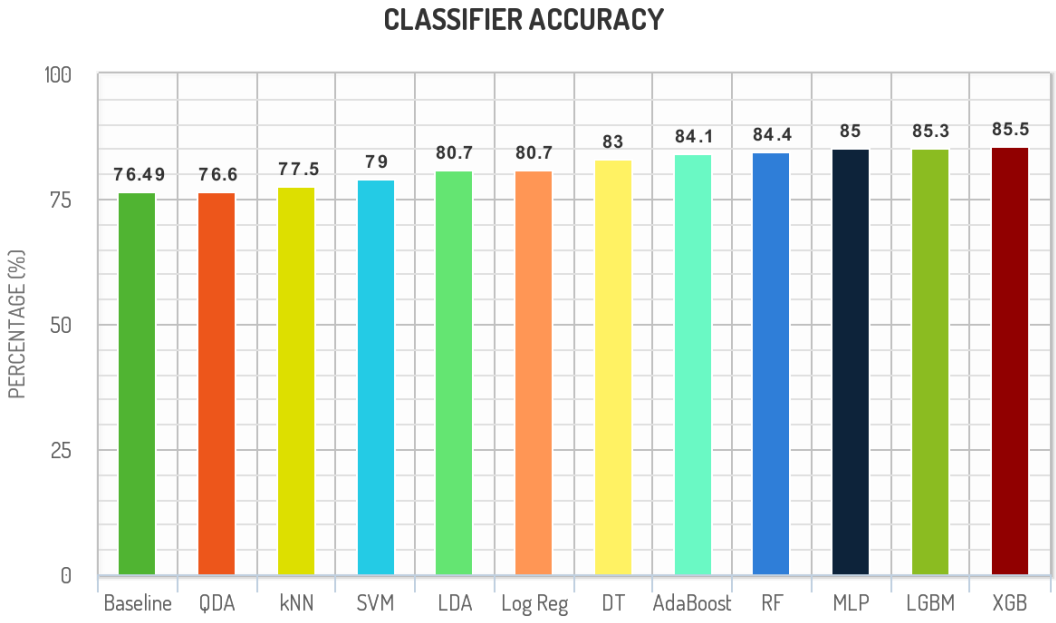
5.3.4 Αποτελέσματα IV

Εξετάζουμε το σύνολο δεδομένων Dataset IV που περιέχει ως πληροφορία τα στοιχεία των πτήσεων με προορισμό το αεροδρόμιο JFK και τα μετεωρολογικά δεδομένα σε αυτό κατά την ώρα της προσγείωσης. Περιλαμβάνει επιπλέον τον όγκο όλων των πτήσεων ημερήσια στο NAS και τον καιρό στο αεροδρόμιο αναχώρησης καθώς και δεδομένα που αφορούν εάν το αεροσκάφος που εκτελεί το δρομολόγια έχει καθυστερήσει σε προηγούμενη πτήση και πόσο. Μετά το πέρας της εκπαίδευσης και της βελτιστοποίησης υπερ-παραμέτρων, τα μοντέλα μας στο σύνολο Dataset IV έδωσαν τα ακόλουθα αποτελέσματα:

Μοντέλο	**Precision**	**Recall**	**F1**	**Test Accuracy**	**Train Accuracy**
kNN	0.75	0.77	0.75	0.775	0.778
Random Forest	0.84	0.84	0.82	0.844	0.888
Linear Discriminant Analysis	0.79	0.81	0.79	0.807	0.807
Logistic Regression	0.79	0.81	0.78	0.807	0.808
AdaBoost Classifier	0.84	0.84	0.82	0.841	0.840

Quadratic Discriminant Analysis	0.73	0.77	0.74	0.766	0.768
Decision Tree	0.82	0.83	0.81	0.830	0.847
Support Vector Machines	0.77	0.79	0.78	0.790	0.792
Light GBM	0.84	0.85	0.84	0.853	0.896
XGBoost	**0.85**	**0.85**	**0.84**	**0.855**	**0.982**
MLP	0.84	0.85	0.83	0.850	0.901

Πίνακας 5: *Αποτελέσματα Εκπαίδευσης στο Dataset IV*



Σχήμα 39: *Γραφική αναπαράσταση ορθότητας ταξινομητών για το σύνολο ελέγχου του Dataset IV*

Στα τελικά δεδομένα εκπαίδευσης που κατασκευάσαμε μπορούμε να καταλάβουμε πόσο καθοριστικό ρόλο στην πρόβλεψη καθυστέρησης μιας πτήσης παίζει η πληροφορία προηγούμενης αργοπορίας. Στους περισσότερους ταξινομητές τα ποσοστά έχουν πλέον "εκτοξευθεί", ενώ παράλληλα για πρώτη φορά η επίδοση όλων των μοντέλων είναι πάνω από το Baseline. Συνολικά χειρότερος είναι ξανά ο QDA ο οποίος με 76.6% accuracy φαίνεται πως δεν είναι χρήσιμος στα δεδομένα μας αφού η διαφορά από το baseline είναι αμελητέα. Ο KNN

παρουσιάζει μια πολύ μικρή αύξηση. Δυνατότητα εκπαίδευσης παρουσιάζει πλέον και ο SVM ο οποίος για πρώτη φορά δεν ταξινομεί όλα τα δεδομένα στην κλάση “On-Time”, αλλά αποδίδει αρκετά καλά (79%). Παρατηρούμε ότι σε όλα τα πειράματα ο ταξινομητής Logistic Regression είναι πολύ κοντά στον Linear Discriminant Analysis, μοτίβο που συνεχίζεται και τώρα με το Accuracy τους στο 80.7%. Ο αλγόριθμος MLP παρουσίασε τεράστια άνοδο στην επίδοση (85%), κάτι που υποδεικνύει ότι ένα τεχνητό νευρωνικό δίκτυο μπορεί αν δώσει αρκετά αξιόπιστα αποτελέσματα σε ένα τέτοιο πρόβλημα. Αξιοσημείωτη είναι η επίδοση των αλγόριθμων Random Forest και Decision Trees οι οποίοι είναι από τους καλύτερους ταξινομητές για το συγκεκριμένο πρόβλημα, με ορθότητα 84.4%. Οι αλγόριθμοι ενίσχυσης, AdaBoost, Light GBM και XGBoost έχουν πολύ ψηλές επιδόσεις με 84,1%, 85,3% και 85,5% αντίστοιχα.

Όπως και στα προηγούμενα πειράματα, έτσι και στο τελικό και σημαντικότερο dataset, ο ταξινομητής **XGBoost** είχε την καλύτερη επίδοση, φτάνοντας μέχρι και το εντυπωσιακό **85.5%**, 9% πάνω από το baseline. Σημειώνεται ότι ο αλγόριθμος τείνει ξεκάθαρα να υπερπροσαρμόζεται πάνω στα δεδομένα εκπαίδευσης αφού στα πειράματα IV, το ποσοστό accuracy του στα δεδομένα του συνόλου εκπαίδευσης έφτασε το 98.2%. Αξίζει να σημειωθεί ότι οι δύο καλύτεροι ταξινομητές βρίσκονται πολύ κοντά σε επίδοση στα δεδομένα ελέγχου, όμως ο Light GBM πετυχαίνει accuracy 89.6% στα δεδομένα εκπαίδευσης που είναι πολύ χαμηλότερο από αυτό του XGBoost. Αυτό σημαίνει ότι δεν παρουσιάζεται έντονή υπερπροσαρμογή, γεγονός που ίσως κάνει τον Light GBM μια καλύτερη επιλογή λόγω καλύτερης γενίκευσης του μοντέλου σε νέα δεδομένα.

Ακολουθεί πίνακας με τις βέλτιστες τιμές υπερ-παραμέτρων στα τελικά πειράματα.

Μοντέλο	**Υπερ-Παράμετροι**
kNN	'n_neighbors': 13
Random Forest	'bootstrap': False, 'max_depth': 12, 'max_features': 'log2', 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 44
Logistic Regression	'C': 4.281332398719396, 'max_iter': 100, 'penalty': 'l1', 'solver': 'liblinear'

AdaBoost Classifier	'algorithm': 'SAMME.R', 'learning_rate': 0.3, 'n_estimators': 90
Decision Tree	'criterion': 'entropy', 'max_depth': 9, 'max_features': 'log2', 'min_samples_leaf': 1, 'min_samples_split': 2, 'splitter': 'best'
Support Vector Machines	'pca__n_components': 35, 'kernel': 'sigmoid', 'C': 0.1, 'gamma': 0.001
Light GBM	'colsample_bytree': 0.77616281347616, 'min_child_samples': 159, 'min_child_weight': 0.001, 'num_leaves: 38, 'reg_alpha': 7, 'reg_lambda': 50, 'subsample': 0.3129741395033088
XGBoost	'colsample_bytree': 0.3, 'eta': 0.05, 'gamma': 0.4, 'learning_rate': 0.1, 'max_depth': 10, 'min_child_weight': 1, 'subsample': 1
MLP	'activation': 'tanh', 'alpha': 0.0005, 'hidden_layer_sizes': (10, 20), 'learning_rate': 'adaptive', 'solver': 'lbfgs', 'pca__n_components': 21

Πίνακας 6: *Βέλτιστες τιμές Υπέρ-Παραμέτρων Μοντέλων*

Κεφάλαιο 6

ΕΠΙΛΟΓΟΣ

6.1 Συμπεράσματα

Στην παρούσα εργασία χρησιμοποιήθηκαν τεχνικές Επιβλεπόμενης Μάθησης με σκοπό την πρόβλεψη καθυστέρησης πτήσεων με την ταξινόμηση τους σε δύο κατηγορίες, τις "On-Time" και "Delayed". Σκοπός ήταν η μελέτη τόσο του ενδεχομένου επιτυχούς πρόβλεψης των καθυστερήσεων χρήσει τεχνικών Μηχανικής Μάθησης, όσο και η διερεύνηση των παραγόντων εκείνων, που συμβάλλουν σημαντικότερα καθορίζοντας την επιτυχία της πρόβλεψης.

Μετά τη διεξαγωγή πειραμάτων με διάφορα μοντέλα ταξινομητών, οι οποίοι χρησιμοποιούν ποικίλες τεχνικές για την επίλυση του διερευνώμενου προβλήματος, συμπεράναμε ότι οι καλύτεροι αλγόριθμοι για την αντιμετώπιση του, είναι αλγόριθμοι Ensemble και συγκεκριμένα αυτοί που χρησιμοποιούν την τεχνική Ενίσχυσης Κλίσης. Οι τρεις αλγόριθμοι ενίσχυσης οι οποίοι εξετάστηκαν στα πλαίσια της πειραματικής διαδικασίας ήταν οι AdaBoost, Light GBM και XGBoost, με τον τελευταίο να αποδεικνύεται καλύτερο/αποδοτικότερο για το πρόβλημα μας.

Πιο συγκεκριμένα, ο αλγόριθμος **XGBoost** αποδείχτηκε ο ιδανικότερος για την πρόβλεψη του κατά πόσο μία πτήση θα καθυστερήσει και στις 4 συνολικά εκδοχές του προβλήματος. Στο τελικό στάδιο κατάφερε να επιτύχει επίδοση **85.5%** στην πρόβλεψη καθυστερήσεων, ένα αρκετά σημαντικό ποσοστό. Αξίζει να σημειωθεί ότι στον αλγόριθμο αυτό παρατηρήθηκε το φαινόμενο της Υπερπροσαρμογής στο σύνολο εκπαίδευσης σε ποσοστό που πλησίασε το 100% προς το τέλος των πειραμάτων. Αξιοσημείωτο είναι επίσης ότι ο δεύτερος καλύτερος αλγόριθμος, Light GBM πλησίασε κατά πολύ στην επίδοση του πρώτου, (μόλις 0.2% διαφορά), με επίδοση accuracy 85.3%. Ωστόσο στον συγκεκριμένο, το πρόβλημα της Υπερπροσαρμογής δεν ήταν τόσο σοβαρό, κάτι που ενδεχομένως να οδηγεί στην επιλογή του ως ακόμα καλύτερου ταξινομητή για γενίκευση του προβλήματος.

Άξιό αναφοράς είναι ότι το Τεχνητό Νευρωνικό Δίκτυο που κατασκευάστηκε για να αντιμετωπίσει το πρόβλημα έφτασε πολύ κοντά στις επιδόσεις των αλγόριθμων ενίσχυσης κλίσης και παρέχει μια καλή εναλλακτική μέθοδο.

Δείχθηκε ακόμα, ότι με ενσωμάτωση στα δεδομένα των κατάλληλων χαρακτηριστικών των καιρικών συνθηκών στα αεροδρόμια προέκυψε βελτίωση στην ικανότητα ταξινόμησης των πτήσεων ως προς την καθυστέρηση. Αναμφίβολα όμως, ο κυρίαρχος παράγοντας αποδείχτηκε πως είναι η καθυστέρηση του αεροσκάφους σε προηγούμενο δρομολόγιο και η διάδοση αυτής της καθυστέρησης στο Εθνικό Αεροπορικό Σύστημα ΗΠΑ. Το συγκεκριμένο πρόβλημα έχει μελετηθεί προηγουμένως και έχουν γίνει προσπάθειες από ερευνητές να εκτιμηθεί η διάδοση της καθυστέρησης μέσα σε αεροπορικό δίκτυο (Chen 2012, Wang, Schaefer 2003, Xu, Donohue 2005) [8], [39], [7].

Εν γένει, τα αποτελέσματα των πειραμάτων ήταν επιτυχημένα και έδειξαν ότι με τη μελέτη των κατάλληλων χαρακτηριστικών και τη διαθεσιμότητα της πληροφορίας μπορούν να προβλεφθούν ικανοποιητικώς αξιόπιστα οι καθυστερήσεις πτήσεων με τεχνικές Επιβλεπόμενης Μάθησης.

6.2 Μελλοντικές Επεκτάσεις

Η παρούσα εργασία θα μπορούσε να επεκταθεί σε διάφορες κατευθύνσεις. Αρχικά θα μπορούσε πέραν του αεροδρομίου JFK να μελετηθεί και άλλο αεροδρόμιο ως αυτό της τελικής άφιξης. Επιπλέον, υπάρχει η δυνατότητα μελέτης κι άλλων αεροδρομίων αναχώρησης διευρύνοντας το δίκτυο των αεροδρομίων που έχουν μελετηθεί. Το πλήθος των αεροδρομίων που υπάρχουν στις ΗΠΑ είναι πολύ μεγάλο και η μελέτη ενός δικτύου τέτοιου μεγέθους μπορεί να γίνει σταδιακά ή τμηματικά. Το πρόβλημα ως όλο θα μπορούσε να αποτελέσει ένα τεράστιο εγχείρημα που ίσως να ενδιέφερε ακόμη και την ίδια την κυβέρνηση των ΗΠΑ.

Ακόμα πιο ενδιαφέρον ερευνητικά και το επόμενο λογικό βήμα σε μια τέτοια εργασία είναι η πρόβλεψη της καθυστέρησης της πτήσης επιπλέον της δυαδικής ταξινόμησης της σε καθυστερημένη ή μη. Κάτι τέτοιο μπορεί να γίνει εφικτό, χωρίζοντας τις καθυστερήσεις σε κατηγορίες, π.χ. σε δεκαπεντάλεπτα παράθυρα, όπου η πρόβλεψη μας θα αποτελεί τα λεπτά της καθυστέρησης και το δεκαπεντάλεπτο παράθυρο στο οποίο θα βρεθεί η πτήση. Ως εκ τούτου, θα μπορεί να γίνει η χρήσιμη διάκριση και παροχή της πληροφορίας στον χρήστη, σε πτήσεις που θα καθυστερήσουν λίγο ή σε πτήσεις που θα καθυστερήσουν κάποιες ώρες.

Βιβλιογραφία

[1] M. Ball *et al.*, “Total delay impact study : a comprehensive assessment of the costs and impacts of flight delay in the United States,” Oct. 2010, [Online]. Available: https://rosap.ntl.bts.gov/view/dot/6234

[2] I. Stamos, E. Mitsakis, J. M. Salanova, and G. Aifadopoulou, “Impact assessment of extreme weather events on transport networks: A data-driven approach,” *Transp. Res. Part Transp. Environ.*, vol. 34, pp. 168–178, Jan. 2015, doi: 10.1016/j.trd.2014.11.002.

[3] Bureau of Transportation Statistics, “On-Time performance - flight delay at a Glance,” 2021. https://www.transtats.bts.gov/HomeDrillChart.asp (accessed Jul. 05, 2021).

[4] W. E. Bendinelli, H. F. A. J. Bettini, and A. V. M. Oliveira, “Airline delays, congestion internalization and non-price spillover effects of low cost carrier entry,” *Transp. Res. Part Policy Pract.*, vol. 85, pp. 39–52, Mar. 2016, doi: 10.1016/j.tra.2016.01.001.

[5] M. Lukacs, “Cost of Delay Estimates 2019,” *Fed. Aviat. Adm.*, Jul. 2020.

[6] A. Kondo, “Delay Propagation and Multiplier,” Transportation Research Forum, 207266, Mar. 2010. Accessed: Jul. 05, 2021. [Online]. Available: https://ideas.repec.org/p/ags/ndtr10/207266.html

[7] N. Xu, G. Donohue, K. B. Laskey, and C.-H. Chen, “Estimation of Delay Propagation in the National Aviation System Using Bayesian Networks,” 2005. https://www.semanticscholar.org/paper/Estimation-of-Delay-Propagation-in-the-National-Xu-Donohue/9c5a7c726315387acb5ff5cf8e849524046cd207 (accessed Jul. 05, 2021).

[8] P. T. R. Wang, L. A. Schaefer, and L. A. Wojcik, “Flight connections and their impacts on delay propagation,” in *Digital Avionics Systems Conference, 2003. DASC ’03. The 22nd*, Oct. 2003, vol. 1, p. 5.B.4-5.1-9 vol.1. doi: 10.1109/DASC.2003.1245858.

[9] G. Kulesa, “Weather and Aviation: How does Weather Affect the Safety and Operations of Airports and Aviation, and how does FAA work to manage Weather-Related Effects?,” presented at the The Potential Impacts of Climate Change on TransportationUS Department of Transportation Center for Climate Change and Environmental Forecasting; US Environmental Protection Agency; US Department of Energy; and US Global Change Research Program, 2003. Accessed: Jul. 06, 2021. [Online]. Available: https://trid.trb.org/view/663829

[10] S. Borsky and C. Unterberger, “Bad weather and flight delays: The impact of sudden and slow onset weather events,” *Econ. Transp.*, vol. 18, pp. 10–26, Jun. 2019, doi: 10.1016/j.ecotra.2019.02.002.

[11] C. M. Bishop, *Pattern recognition and machine learning*, 13. (corrected at 8th printing 2009). New York: Springer, 2009.

[12] E. Alpaydin, *Introduction to machine learning*. Cambridge, Mass: MIT Press, 2004.

[13] S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach*, 2nd ed. Upper Saddle River, N.J: Prentice Hall/Pearson Education, 2003.

[14] B. W. Silverman and M. C. Jones, “E. Fix and J.L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951),” *Int. Stat. Rev. Rev. Int. Stat.*, vol. 57, no. 3, pp. 233–238, 1989, doi: 10.2307/1403796.

[15] D. R. Cox and E. J. Snell, *Analysis of binary data*, 2nd ed. London ; New York: Chapman and Hall, 1989.

[16] S. Sperandei, “Understanding logistic regression analysis,” *Biochem. Medica*, pp. 12–18, 2014, doi: 10.11613/BM.2014.003.

[17] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*, Third edition. Hoboken, New Jersey: Wiley, 2013.

[18] R. A. Fisher, “The Use of Multiple Measurements in Taxonomic Problems,” *Ann. Eugen.*, vol. 7, no. 2, pp. 179–188, 1936, doi: 10.1111/j.1469-1809.1936.tb02137.x.

[19] J. R. Quinlan, “Induction of decision trees,” *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1007/BF00116251.

[20] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.

[21] T. K. Ho, “Random decision forests,” in *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*, USA, Aug. 1995, p. 278.

[22] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[23] C. Zhang and Y. Ma, Eds., *Ensemble Machine Learning*. New York: Springer, 2012.

[24] M. Kearns and L. G. Valiant, “Crytographic limitations on learning Boolean formulae and finite automata,” in *Proceedings of the twenty-first annual ACM symposium on Theory of computing*, New York, NY, USA, Feb. 1989, pp. 433–444. doi: 10.1145/73007.73049.

[25] R. E. Schapire, “The strength of weak learnability,” *Mach. Learn.*, vol. 5, no. 2, pp. 197–227, Jun. 1990, doi: 10.1007/BF00116037.

[26] R. E. Schapire, “The Boosting Approach to Machine Learning: An Overview,” in *Nonlinear Estimation and Classification*, D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, and B. Yu, Eds. New York, NY: Springer, 2003, pp. 149–171. doi: 10.1007/978-0-387-21579-2_9.

[27] Y. Freund and R. E. Schapire, “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting,” *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997, doi: 10.1006/jcss.1997.1504.

[28] J. H. Friedman, “Greedy function approximation: A gradient boosting machine.,” *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/aos/1013203451.

[29] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[30] G. Ke *et al.*, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” Dec. 2017. [Online]. Available: https://www.microsoft.com/en-us/research/publication/lightgbm-a-highly-efficient-gradient-boosting-decision-tree/

[31] S. S. Haykin, *Neural networks: a comprehensive foundation*, 2nd ed. Upper Saddle River, N.J: Prentice Hall, 1999.

[32] F. Rosenblatt, “The perceptron: A probabilistic model for information storage and organization in the brain.,” *Psychol. Rev.*, vol. 65, no. 6, pp. 386–408, 1958, doi: 10.1037/h0042519.

[33] M. Minsky and S. A. Papert, *Perceptrons: an introduction to computational geometry*, 2. print. with corr. Cambridge/Mass.: The MIT Press, 1972.

[34] P. J. Werbos, *The roots of backpropagation: from ordered derivatives to neural networks and political forecasting*. New York: Wiley, 1994.

[35] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986, doi: 10.1038/323533a0.

[36] National Climatic Data Center, NESDIS, NOAA, U.S. Department of Commerce, “NOAA National Centers for Environmental Information (2001): Global Surface Hourly.”

National Centers for Environmental Information, NESDIS, NOAA, U.S. Department of Commerce.

[37] J. E. Jackson, *A user's guide to principal components*. Hoboken, N.J: Wiley-Interscience, 2003.

[38] M. Richardson, "Principal component analysis," 2009.

[39] D. M. Hawkins, "The Problem of Overfitting," *J. Chem. Inf. Comput. Sci.*, vol. 44, no. 1, pp. 1–12, Jan. 2004, doi: 10.1021/ci0342472.

[40] K. B. Laskey, N. Xu, and C.-H. Chen, "Propagation of Delays in the National Airspace System," *ArXiv12066859 Cs*, Jun. 2012, Accessed: Jul. 05, 2021. [Online]. Available: http://arxiv.org/abs/1206.6859

Παράρτημα

Ελληνικοί όροι

ΜΜ: Μηχανική Μάθηση

ΤΝΔ: Τεχνητό Νευρωνικό Δίκτυο

Αγγλικοί όροι

NAS: National Airspace System

BTS: Bureau of Transportation Statistics

FAA: Federal Aviation Administration

NTSB: National Transportation Safety Board

NOAA: National Oceanic and Atmospheric Administration

KNN: K-Nearest Neighbors

MLE: Maximum Likelihood Estimation

OLS: Ordinary Least Squares

LDA: Linear Discriminant Analysis

QDA: Quadratic Discriminant Analysis

RF: Random Forest

SVM: Support Vector Machines

MLP: Multi-Layer Perceptron

MSE: Mean Squared Error

MAE: Mean Absolute Error

UTC: Coordinated Universal Time

PCA: Principal Component Analysis